



**NUMERICAL PROCEDURES FOR  
ANALYZING DYNAMICAL PROCESSES**

**FINAL REPORT**



**Project Period: October 1, 1990 – February 29, 1992**

**Celso Grebogi  
Edward Ott  
James A. Yorke**

**University of Maryland  
College Park, MD 20742-3511**

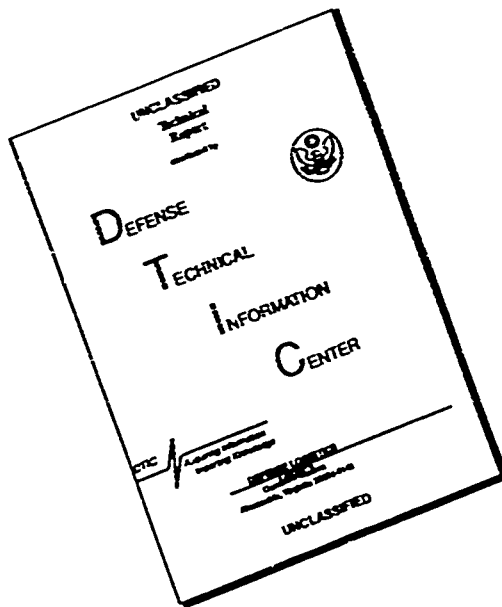
**92-07966**



**92 3 30 047**

**Prepared for the Office of Naval Research (DARPA) under grant  
number N00014-88-K-0657**

# DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

**OFFICE OF NAVAL RESEARCH  
PUBLICATIONS / PATENTS / PRESENTATIONS / HONORS  
FOR  
1 OCTOBER 1990 THROUGH 29 FEBRUARY 1992**

<b>CONTRACT:</b>	N00014-88-K-0657
<b>R&amp;T NO.:</b>	b41u001- - -04
<b>TITLE OF CONTRACT:</b>	Numerical Procedures for Analyzing Dynamical Processes
<b>NAME(S) OF PRINCIPAL INVESTIGATOR(S):</b>	Celso Grebogi Edward Ott James A. Yorke
<b>NAME OF ORGANIZATION:</b>	University of Maryland
<b>ADDRESS OF ORGANIZATION:</b>	Laboratory for Plasma Research College Park, Maryland 20742-3511

Reproduction in whole, or in part, is permitted for any purpose of the United States Government.


This document has been approved for public release and sale; its distribution is unlimited.

## INTRODUCTION

The following report summarizes our activities under the Office of Naval Research (DARPA) Contract No. N0014-88-K-0657. We have organized this report under the following five categories:

- I. Deliverables: computer tape and disk with instructions, and summary of accomplishments related to the proposed projects.
- II. List of publications for the period of this report.
- III. Appended respective reprints and preprints.

Accession For	
NTIS GR&I	<input checked="checked" type="checkbox"/>
DIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	23



## **I. Deliverables: Computer Tape and Disk with Instructions and Summary of Accomplishments Related to the Proposed Projects**

### **PROJECT I**

*Deliverable:* We are delivering a tape with a software package for UNIX workstations with documentation for analyzing low dimensional dynamical behavior from time series. In particular, the Lyapunov exponent code will, together with the dimension code, permit the user to distinguish between periodic, chaotic, and random processes. "Random processes" here means behavior whose dimension is too high to compute. The code computes the information dimension of the time series. We are also including in the same tape a noise-reduction code with documentation.

#### *Summary of Project 1: Nonlinear Noise Filtering of Experimental Data from Chaotic Processes*

Many attempts have been made to apply ideas from dynamical systems to the analysis of experimental data including estimates of attractor dimension and measurement of Lyapunov exponents. An essential problem is that noise often complicates the analysis. For example, noise obscures the fractal structure of the attractor, so that estimates of the attractor dimension can be difficult to obtain. Various methods have been proposed to estimate the noise levels in the data, and these are useful for determining the smallest scales at which dimension measurements are feasible. However, up until now no systematic method has been developed for noise reduction.

We have developed a method which we believe is a potential breakthrough in the analysis of experimental data. Typically, attractors are reconstructed from a scalar time series of experimental data using time delays. Conventional signal filtering techniques are not useful in this case, because they examine only portions of the signal which are close in time. We examine points on an attractor which are close in

phase space; the corresponding parts of the original signal in general are far apart in time.

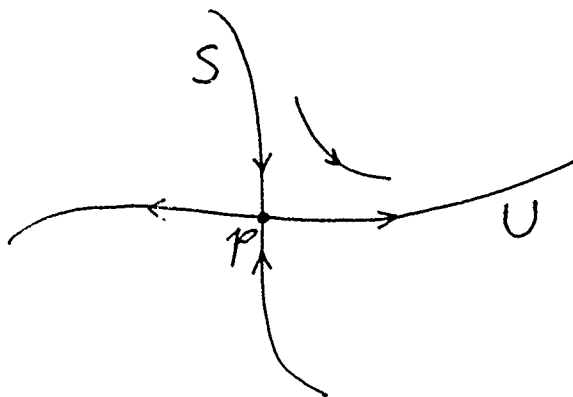
Our method is a linearization technique which uses the dynamics of the reconstructed attractor to estimate and correct errors in the trajectories. The method relies on the assumption that in a small neighborhood about a point on the trajectory, the dynamics on the attractor is nearly linear. In other words, given a point  $\mathbf{x}_i$  on the attractor, its image is  $\mathbf{x}_{i+1} = \mathbf{f}(\mathbf{x}_i)$  for some nonlinear, unknown function  $\mathbf{f}$ . We assume that it is possible to find a matrix  $\mathbf{A}$  and a vector  $\mathbf{b}$  such that  $\mathbf{x}_{i+1} \approx \mathbf{A}\mathbf{x}_i + \mathbf{b}$ . The method has two steps: first, to compute the matrices  $\mathbf{A}$  and vectors  $\mathbf{b}$  for each point on the trajectory, and second, to find a new trajectory near the original one which best satisfies the linear approximation. We believe that a reliable procedure like the one outlined above will be invaluable for the analysis of experimental data.

## PROJECT 2

*Deliverable:* We are delivering a disk containing a Dynamics code (for IBM compatible PCs and for UNIX/X window workstations) with a Manual for computing and evaluating dynamical processes. The source code contains 20,000 lines of code. In particular, the program will compute stable and unstable manifolds as described below.

*Summary of Project 2: A fast Reliable Method for the Numerical Computation of Stable Manifolds of Chaotic Processes*

Saddle points often play a crucial role in the dynamics of a particular map  $f$ . A schematic illustration of a saddle point in two dimensions is given in the following figure:



Because  $p$  is an unstable fixed point, any point  $p'$  eventually moves away from  $p$  as  $f$  is iterated, even though  $f(p) = p$ . For example, in Fig. 1, initial conditions slightly to the right of the curve labeled  $S$  move toward  $p$  for a few iterates, then are repelled to the right thereafter, eventually approaching the curve  $U$ . Initial conditions slightly to the left of  $S$  will move close to  $p$ , then off to the left. The curve  $S$  is the *stable manifold* of  $p$ : it is the set of initial conditions which are attracted to  $p$ . The curve  $U$  is the *unstable manifold*. If  $f$  is invertible then  $U$  is the stable manifold of  $p$  for the inverse map  $f^{-1}$ . More generally,  $U$  is the set of points whose preimages tend to  $p$ .

In many cases, the stable and unstable manifolds wind around in complicated ways. Because the manifolds are intertwined so closely, initial conditions can approach and be repelled from the saddle point repeatedly, leading to complex behavior. Stable manifolds of fixed points often form part of the boundary between two basins of attraction. In this case, the structure of the stable manifold determines how sensitive the system is to small errors in measuring an initial condition. In addition, it is often important to know whether the stable and unstable manifolds cross at a point other than the saddle point  $p$ . Such *homoclinic intersections* are often of interest, especially in cases where the map depends on a parameter. Hence, a knowledge of the structure of the stable and unstable manifolds is essential to understanding the dynamics. We have developed efficient, reliable numerical methods to calculate them.

## II. LIST OF PUBLICATIONS FOR THE PERIOD OF THIS REPORT

1. "Noise Reduction: Finding the Simplest Dynamical System Consistent with the Data," E. J. Kostelich and J. A. Yorke, *Physica* **41D**, 183 (1990).
2. "Shadowing of Physical Trajectories in Chaotic Dynamics: Containment and Refinement," C. Grebogi, S. M. Hammel, J. A. Yorke, and T. Sauer, *Phys. Rev. Lett.* **65**, 1527 (1990).
3. Antimonotonicity: Concurrent Creation and Annihilation of Periodic Orbits," *Bulletin AMS* **23**, 469 (1990).
4. "Chaotic Scattering in Several Dimensions," Q. Chen, M. Ding, and E. Ott, *Phys. Lett.* **145A**, 93 (1990).
5. "Cross-sections of Chaotic Attractors," Q. Chen and E. Ott, *Phys. Lett.* **147A**, 450 (1990).
6. "Rigorous Verification of Trajectories for the Computer Simulation of Dynamical Systems," T. Sauer and J. A. Yorke, *Nonlinearity* **4**, 961 (1991).
7. "Analysis of a Procedure for Finding Numerical Trajectories Close to Chaotic Saddle Hyperbolic Sets," H. E. Nusse and J. A. Yorke, *Ergod. Th. & Dynam. Sys.* **11**, 189 (1991).
8. "Embedology," T. Sauer, J. A. Yorke, and M. Casdagli, *J. Stat. Phys.* **65**, 579 (1991).
9. "A Numerical Procedure for Finding Accessible Trajectories of Basin Boundaries," H. E. Nusse and J. A. Yorke, *Nonlinearity* **4**, 1183 (1991).
10. "Calculating Topological Entropies of Chaotic Dynamical Systems," Q. Chen, E. Ott, and L. Hurd, *Phys. Lett.* **156A**, 48 (1991).
11. "On the Tendency Toward Ergodicity with Increasing Number of Degrees of Freedom in Hamiltonian," L. Hurd, C. Grebogi, and E. Ott, submitted for publication.

12. "Metamorphoses: Sudden Jumps in Basin Boundaries," K. T. Alligood, L. Tedeschini-Lalli, and J. A. Yorke, submitted for publication.
13. "Accessible Saddles on Fractal Basin Boundaries," K. T. Alligood and J. A. Yorke, submitted for publication.
14. "The Analysis of Experimental Data Using Time-Delay Embedding Methods," E. J. Kostelich and J. A. Yorke, submitted for publication.
15. "When Cantor Sets Intersect Thickly," B. R. Hunt, I. Kan, and J. A. Yorke, submitted for publication.
16. "Border-Collision Bifurcations Including 'Period Two to Period Three' for Piecewise Smooth Systems," H. E. Nusse and J. A. Yorke, submitted for publication.

**III. APPENDED RESPECTIVE REPRINTS AND  
PREPRINTS**

## NOISE REDUCTION: FINDING THE SIMPLEST DYNAMICAL SYSTEM CONSISTENT WITH THE DATA

Eric J. KOSTELICH<sup>a,b,1</sup> and James A. YORKE<sup>a,c</sup>

<sup>a</sup>*Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA*

<sup>b</sup>*Center for Nonlinear Dynamics, Department of Physics, University of Texas, Austin, TX 78712, USA*

<sup>c</sup>*Department of Mathematics, University of Maryland, College Park, MD 20742, USA*

Received 3 March 1989

Revised manuscript received 11 October 1989

Accepted 18 October 1989

Communicated by R. Westervelt

A novel method is described for noise reduction in chaotic experimental data whose dynamics are low dimensional. In addition, we show how the approach allows experimentalists to use many of the same techniques that have been essential for the analysis of nonlinear systems of ordinary differential equations and difference equations.

### 1. Introduction

Numerical computation and computer graphics have been essential tools for investigating the behavior of nonlinear maps and differential equations. The pioneering work of Lorenz [25] was made possible by numerical integration on a computer, allowing him to take nearby pairs of initial conditions and compare the trajectories. Hénon [19] discovered the complex dynamics of his celebrated quadratic map with the aid of a programmable calculator. A variety of classical and modern techniques has been exploited to find periodic orbits, their stable and unstable manifolds [14], basins of attraction [26], fractal dimension [27], and Lyapunov exponents [10, 31, 37]. In some cases, numerical methods can establish rigorously the existence of initial conditions whose trajectories have essentially the same intricate structure that one sees on a computer screen [18].

Until recently, experimentalists have not been able to apply most of these methods to the analysis of experimental data, since they do not in general have explicit equations to model the behavior of their apparatus. In cases where it is possible to find accurate models of the physical system, quantitative predictions about the behavior of actual experiments are possible [17]. However, all that is available in a typical experiment is the time-dependent output (e.g. voltage) from one or more probes, which is a function of the dynamics.

One fundamental problem in the analysis of experimental data concerns the correspondence between the dynamics that governs the behavior of the apparatus and the discretely sampled time series that comprises the data. Another question is how to minimize the effect of noise. In this paper, we show how the *time delay embedding method*, now commonly used to reconstruct an attractor from experimental data, yields a novel procedure for reducing noise in data whose dynamics can be characterized as low dimensional. Moreover, we

<sup>1</sup>Current address: Department of Mathematics, Arizona State University, Tempe, AZ 85287, USA

show how the approach can be extended to allow experimentalists access to many of the analytical tools mentioned above.

Section 2 reviews the time delay embedding method and some of its applications. Section 3 introduces some of the problems associated with traditional filters and outlines our noise reduction method.

## 2. The time delay embedding method

As stated in section 1, one problem in analyzing experimental data is how to relate the measurements with the dynamics. Before the early 1980's, power spectra were the principal method for analyzing such data. For instance, Fenstermacher et al. [13] relied heavily on power spectra to detect transitions from periodic to weakly turbulent flow between concentric rotating cylinders. However, Fourier analysis alone is inadequate for describing the dynamics.

Other methods also have been used to analyze time series output from dynamical systems. Lorenz [25] used next amplitude maps to describe some features of the dynamics; that is, he plotted  $z_{n+1}$  against  $z_n$ , where  $z_n$  is the  $n$ th relative maximum of the third coordinate of the numerically calculated solution. Such maps are often useful, not only for investigating features of the Lorenz attractor [32], but also for instance in experiments on intermittency in oscillating chemical reactions [30].

In the past decade, the time delay embedding method has come into common use as a way of reconstructing an attractor from a time series of experimental data. In this approach, one supposes that the dynamical behavior is governed by a solution traveling along an attractor<sup>#1</sup> (which is not observable directly). However, one assumes there is a smooth function that maps points on the attractor to real numbers (the experimental

measurements). In the embedding method, one generates a set of  $m$ -dimensional points whose coordinates are values in the time series separated by a constant delay [11]. For example, when  $m = 3$ , the reconstructed attractor is the set of points  $\{x_i = (s_i, s_{i+\tau}, s_{i+2\tau})\}$  where  $\tau$  is the time delay. Takens [34] has shown that under suitable hypotheses, this procedure yields a set whose properties are equivalent to those of the original attractor provided that the embedding dimension  $m$  is large enough.

In principle, the embedding method allows one to study the dynamics in detail. The earliest applications may be called *static* in that the analysis focuses on the geometric properties of the set of points on the reconstructed attractor. For example, phase portraits and Poincaré sections are used in ref. [5] to help determine the transition between quasiperiodic and chaotic flow in a Couette-Taylor experiment. Another important application is the estimation of attractor dimension from experimental data, for which there is a large literature [27]. In addition, various information theoretic notions can be used to find good choices of embedding dimension and time delay [15].

More recent applications of the embedding method are quite different in nature and can be called *dynamic* in that information about the dynamics is stored in the computer for analysis. With each data vector  $x_i$ , one stores the "next" vector, for example,  $x_{i+\delta}$  for some  $\delta > 0$ . This makes it possible to compute a linear approximation of the dynamics in a neighborhood of  $x_i$ , assuming that there is a low-dimensional dynamical system underlying that data<sup>#2</sup>. In particular, a linear approximation provides an estimate of the Jacobian of the map at  $x_i$  [11]. Eckmann et al. [10] use linear maps computed in this way to integrate a set of variational equations and find the positive Lyapunov exponents<sup>#3</sup>.

<sup>#2</sup>This material was first presented by D. Ruelle at a Nobel symposium in 1984.

<sup>#3</sup>Wolf et al. [37] have proposed a different method in which nearby pairs of points are followed to estimate the largest Lyapunov exponent.

<sup>#1</sup>Existing numerical methods require the attractor to be low dimensional.

In fact, the time delay embedding method provides a powerful set of tools for analyzing the dynamics, the breadth of which may not have been realized by Eckmann and Ruelle. In the remainder of this paper, we discuss two novel applications that are possible, specifically:

(1) *Noise reduction.* Since one can approximate the dynamics at each point, it becomes possible to identify and correct inaccuracies in trajectories arising from random errors in the original time series. Numerical evidence suggests that the noise reduction procedure described below improves the accuracy of other analyses, such as Lyapunov exponents and dimension calculations.

(2) *Simplicial approximations.* Linear approximations can be computed at each point on a grid in a neighborhood of the attractor to form a simplicial approximation of the dynamical system. This can be used to locate unstable periodic orbits near the attractor.

We consider noise reduction in section 3.

### 3. Noise reduction

The ability to extract information from time-varying signals is limited by the presence of noise. Recent experiments to study the transition to turbulence in systems far from equilibrium, like those by Fenstermacher et al. [13], Behringer and Ahlers [2], and Libchaber et al. [24], succeeded largely because of instrumentation that enabled them to quantify and reduce the noise. However, it is often expensive and time consuming to redesign experimental apparatus to improve the signal to noise ratio.

An important question, therefore, is how the experimental data can be filtered or otherwise preprocessed before it is analyzed further. One common approach is to use Fourier analysis: one might model the noise as a collection of high-frequency components and subtract them from a power spectrum (or Fourier transform) of the input data. The transform can be inverted to yield a

new time series with some of the high-frequency components removed. This is the basic idea behind Wiener and other bandpass filters [29].

However, as noted previously, power spectral analysis is insufficient to characterize the dynamics when the data are chaotic. Since the power spectrum of a low-dimensional chaotic signal resembles that of a noisy one, the suppression of certain frequencies can alter the dynamics of the filtered output signal. Badii et al. [1] have shown that a simple low-pass filter effectively introduces an extra Lyapunov exponent that depends on the cutoff frequency. If the cutoff frequency is sufficiently low, then the filter can increase the fractal dimension of the reconstructed attractor. This result also has been confirmed by Mitschke et al. [28] with data from an electronic circuit.

We now consider a different approach and show how the time delay embedding method can be exploited to reduce the noise, at least in cases where the time series can be viewed as a dynamical system with a low-dimensional attractor. Our objective is to use the dynamics to detect and correct errors in trajectories that result from noise. This is done in two steps once an embedding dimension  $m$  and a time delay  $\tau$  have been fixed.

In the first step, we consider the motion of an *ensemble* of points in a small neighborhood of each point on the attractor in order to compute a linear approximation of the dynamics there. In the second step, we use these approximations to consider how well an *individual* trajectory obeys them. That is, we ask how the observed trajectory can be perturbed slightly to yield a new trajectory that satisfies the linear maps better. The trajectory adjustment is done in such a way that a new time series is output whose dynamics are more consistent with those on the phase space attractor.

This approach is fundamentally different from traditional noise reduction methods. Because we consider the motion of points on a phase space attractor, we are using information in the original signal that is not localized in a time or frequency domain. Points that are close in phase space correspond to data that in general are widely and

irregularly spaced in time, due to the sensitive dependence on initial conditions on chaotic attractors. In contrast, Kalman [4] and similar filters examine data that are closely spaced in time: bandpass filters operate in the frequency domain.

#### 4. Eckmann-Ruelle linearization

The discrete sampling of the original signal means that the points on the reconstructed attractor can be treated as iterates of a nonlinear map  $f$  whose exact form is unknown. We assume that  $f$  is nearly linear in a small neighborhood of each attractor point  $x$  and write

$$f(x) \approx Ax + b \equiv L(-)$$

for some  $m \times m$  matrix  $A$  and  $m$ -vector  $b$ . (The matrix  $A$  is the Jacobian of  $f$  at  $x$ .)

This approximation, which we call the *Eckmann-Ruelle linearization* at  $x$ , can be computed with least-squares methods similar to those described in refs. [11, 10]. Given a reference point  $x_{\text{ref}}$ , let  $\{x_i\}_{i=1}^n$  be a collection of the  $n$  points which are closest to  $x_{\text{ref}}$ . With each point  $x_i$ , we store the next point (i.e., the image of  $x_i$ ), denoted  $y_i$ <sup>#4</sup>. The  $k$ th row  $a_k$  of  $A$  and the  $k$ th component  $b_k$  of  $b$  are given by the least-squares solution of the equation

$$y_k = b_k + a_k \cdot x, \quad (1)$$

where  $y_k$  is the  $k$ th component of  $y$  and the dot denotes the dot product. Fig. 1 illustrates the idea<sup>#5</sup>.

<sup>#4</sup>The points  $x_i$  are points on the attractor which are *not* consecutive in time. The subscript  $i$  merely enumerates all the points on the attractor contained within a small distance  $\epsilon$  of  $x_{\text{ref}}$ . In this notation,  $x_i$  and  $y_i$  are consecutive in time.

<sup>#5</sup>Farmer and Sidorowich [12] observe that the Eckmann-Ruelle linearization can be used for prediction. Given a reference point  $x_i$ , find the Eckmann-Ruelle linearization  $A_i x + b_i$ , compute  $x_{i+1} = A_i x_i + b_i$ , and repeat the process to get the predicted trajectory

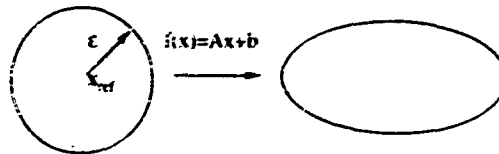


Fig. 1. Schematic diagram for the first stage of the noise reduction method. A collection of points in an  $\epsilon$ -ball about the reference point  $x_{\text{ref}}$  is used to find a linear approximation of the dynamics there.

We mention three difficulties in computing the local linear approximations in the subsections below.

##### 4.1. Ill conditioned least squares

There is a particular problem when one tries to compute solutions to eq. (1) with a finite data set of limited accuracy that has not been addressed in previous papers [10, 31]. Suppose for example that all the points in a neighborhood of  $x_{\text{ref}}$  lie nearly along a single line, i.e., the attractor appears one dimensional within the available resolution. Although it is possible to measure the expansion along the unstable manifold at  $x_{\text{ref}}$ , there are not enough points in other directions to measure the contraction. Hence it is not possible to compute a  $2 \times 2$  Jacobian matrix accurately. Any attempt to do so will result in an estimate of the Jacobian whose elements have large relative errors. This kind of least-squares problem is *ill conditioned*.

The ill conditioning can be avoided by changing coordinates so that the first vector in the new basis points in the unstable direction<sup>#6</sup>. A one-dimensional approximation of the dynamics is computed using the new coordinates; that is, we approximate the dynamics only along the unstable manifold. We recover the matrix  $A$  by changing coordinates back to the original basis.

For example, if we are working in the plane and the unstable direction is the line  $y = x$ , then we rotate the coordinate axes by  $45^\circ$ . The dynamics are approximated by a one-dimensional linear map

<sup>#6</sup>This is done by computing the right singular vectors [9] of the  $n \times m$  matrix whose  $j$ th row is  $x_j$ . The procedure is called *principal component analysis* in the statistical literature.

computed along the line  $y = x$ . Then we rotate back to the original coordinates. (The resulting matrix  $A$  has rank 1 in this example.) This approach substantially enhances the robustness of the numerical procedure.

#### 4.2. Finding nearest neighbors

A second problem is finding an efficient way to locate all of the points closest to a given reference point. The dynamical embedding method imposes stringent requirements on any nearest-neighbor algorithm. The storage overhead for the corresponding data structures must be small, because there are tens of thousands of attractor points. The algorithm must be fast, since there is one nearest-neighbor problem for each linear map to be computed.

We solve this problem by partitioning the phase space into a grid of boxes that is parallel to the coordinate axes. Each coordinate axis is divided into  $B$  intervals. (Fig. 2 illustrates the grid in two dimensions.) Each point on the attractor is assigned a box number according to its coordinates. For example, a point on the plane whose first coordinate falls in the  $j$ th interval (counting from 0) along the  $x$  axis and whose second coordinate falls in the  $k$ th interval along the  $y$  axis is assigned to box number  $kB + j$ . The list of box numbers is sorted, carrying along a pointer to the original data point. Given a reference point  $x_{\text{ref}}$ , its box number is found using the above formula. A binary search in the list of box numbers then locates the address of  $x_{\text{ref}}$  and all the other points

$B^2 - B$	$B^2 - B + 1$	$B^2 - B + 2$		$B^2 - 1$
$B$	$B + 1$	$B + 2$		$2B - 1$
0	1	2		$B - 1$

Fig. 2. Box numbering scheme in two dimensions. The attractor is normalized to fit in the unit square. The bottom row of boxes rests against the  $x$  axis and the leftmost row of boxes against the  $y$  axis.

in the same box number. The search is extended if necessary to adjacent boxes.

Only a crude partition is needed for this algorithm to work efficiently (typically we choose  $B = 40$ ), and the grid is extended only to the first three coordinate axes. When the embedding dimension is larger than three, a preliminary list of nearest neighbors is obtained using only the first three coordinates of each attractor point. The final list is extracted by computing the distances from  $x_{\text{ref}}$  to each point in the preliminary list.

Although there are circumstances where this algorithm can perform poorly (e.g., when most of the attractor points are concentrated in a handful of boxes), the distribution of points on typical attractors is sufficiently uniform that the running time is very fast. Memory use is also efficient: a set of  $N$  attractor points requires  $3N$  storage locations. In contrast, the tree-search algorithm advocated in ref. [12] requires several times more storage (although the lookup time is probably slightly less). Because  $N \approx 10^5$  in typical applications, we believe that the box-grid approach (or some variant) is the most practical. A survey of other nearest-neighbor algorithms is given in ref. [3].

#### 4.3. Errors in variables

There is a potential difficulty in the use of ordinary least squares to compute the linear maps. In the usual statistical problem of fitting a straight line, one has observations  $(x_i, y_i)$  where  $x_i$  is known exactly and  $y_i$  is measured. One assumes that  $y_i = a_0 + a_1 x_i + \epsilon_i$ , where the  $\epsilon_i$  are independent errors drawn from the same normal distribution. (Analogous assumptions hold in the multivariate case.) In the present situation, however, both  $x_i$  and  $y_i$  are measured with error. It can be shown that the ordinary least-squares method produces biased estimates of the parameters  $a_0$  and  $a_1$  in this case [16, 23]. In practice this does not seem to be a serious problem, but statistical procedures to handle this situation (the so-called "errors in variables" methods) may provide

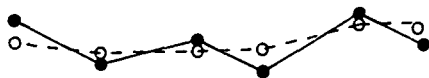


Fig. 3. Schematic diagram of the trajectory adjustment procedure. The trajectory defined by the sequence  $\{x_i\}$  is perturbed to a new trajectory given by  $\{\hat{x}_i\}$  which is more consistent with the dynamics. In this example we show what the perturbed trajectory might look like if the dynamics were approximately horizontal translation to the right.

an alternative approach to noise reduction. We consider this question in the appendix.

### 5. Trajectory adjustment by minimizing self-inconsistency

The Eckmann–Ruelle linearization procedure described above is computed and the resulting maps are stored for a sequence of reference points along a given trajectory (for the results quoted here, the sequence usually contains 24 points). We now consider how to perturb this trajectory so that it is more consistent with the dynamics. The objective is to choose a new sequence of points  $\hat{x}_i$  to minimize the sum of squares

$$\sum w \|\hat{x}_i - x_i\|^2 + \|\hat{x}_i - L_{i-1}(\hat{x}_{i-1})\|^2 + \|\hat{x}_{i+1} - L_i(\hat{x}_i)\|^2, \quad (2)$$

where  $L(x_i) = A_i x_i + b_i$ ,  $w$  is a weighting factor, and the sum runs over all the points along the trajectory<sup>\*7</sup>. Eq. (2) can be solved using least squares. Heuristically, eq. (2) measures the self-inconsistency of the data, assuming that the linear approximations of the dynamics are accurate. See fig. 3. We say the new sequence  $\{\hat{x}_i\}$  is more *self-consistent*.

<sup>\*7</sup>In the results described in this paper, the Eckmann–Ruelle linearization procedure is done using a collection of points within a radius of 1–6% of each reference point, depending on the embedding dimension, the dimension of the attractor, and the number of attractor points. This results in collections of 50–200 points per ball, which gives reasonably accurate map approximations without making the computer program too slow. The weighting factor  $w$  is set to 1.

The trajectory adjustment can be iterated. That is, once a new trajectory  $\hat{x}_i$  has been found, one can replace each  $x_i$  in eq. (2) by  $\hat{x}_i$  and compute a new sequence  $\{\hat{\hat{x}}_i\}$ .

We place an upper limit on the distance a point can move. Points which seem to require especially large adjustments can be flagged and output unchanged. (This may be necessary if the input time series contains large “glitches” or if nonlinearities are significant over small distances in certain regions of the attractor.)

When the input is a time series, we modify the above procedure slightly since we require a time series as output. The trajectory adjustment is done so that changes to the coordinates of  $x_i$  (corresponding to particular time series values) are made consistently for all subsequent points whose coordinates are the same time series values. For example, suppose the time delay is 1 and the embedding dimension is 2. Then trajectories are perturbed so that the second coordinate of the  $i$ th point is the same as the first coordinate of the  $(i+1)$ st point. That is, when  $x_i = (s_i, s_{i+1})$  is moved to the point  $\hat{x}_i = (\hat{s}_i, \hat{s}_{i+1})$ , we require that the first coordinate of  $\hat{x}_{i+1}$  be  $\hat{s}_{i+1}$ .

### 6. Results using experimental data

We note that the attractor need not be chaotic for this noise reduction procedure to be effective. Fig. 4a shows a phase portrait of noisy measurements of wavy vortex flow in a Couette–Taylor experiment [20]. This flow is periodic, so the attractor is a limit cycle (widened into a band because of the noise) and the power spectrum consists of one fundamental frequency and its harmonics above a noise floor. See fig. 4b. Figs. 4c, 4d show the same data after noise reduction. The noise reduction procedure makes the limit cycle much narrower, and the noise floor in the power spectrum is reduced by almost two orders of magnitude. However, no power is subtracted from any of the fundamental frequencies, and in fact some harmonics are revealed which previously were obscured by the noise.

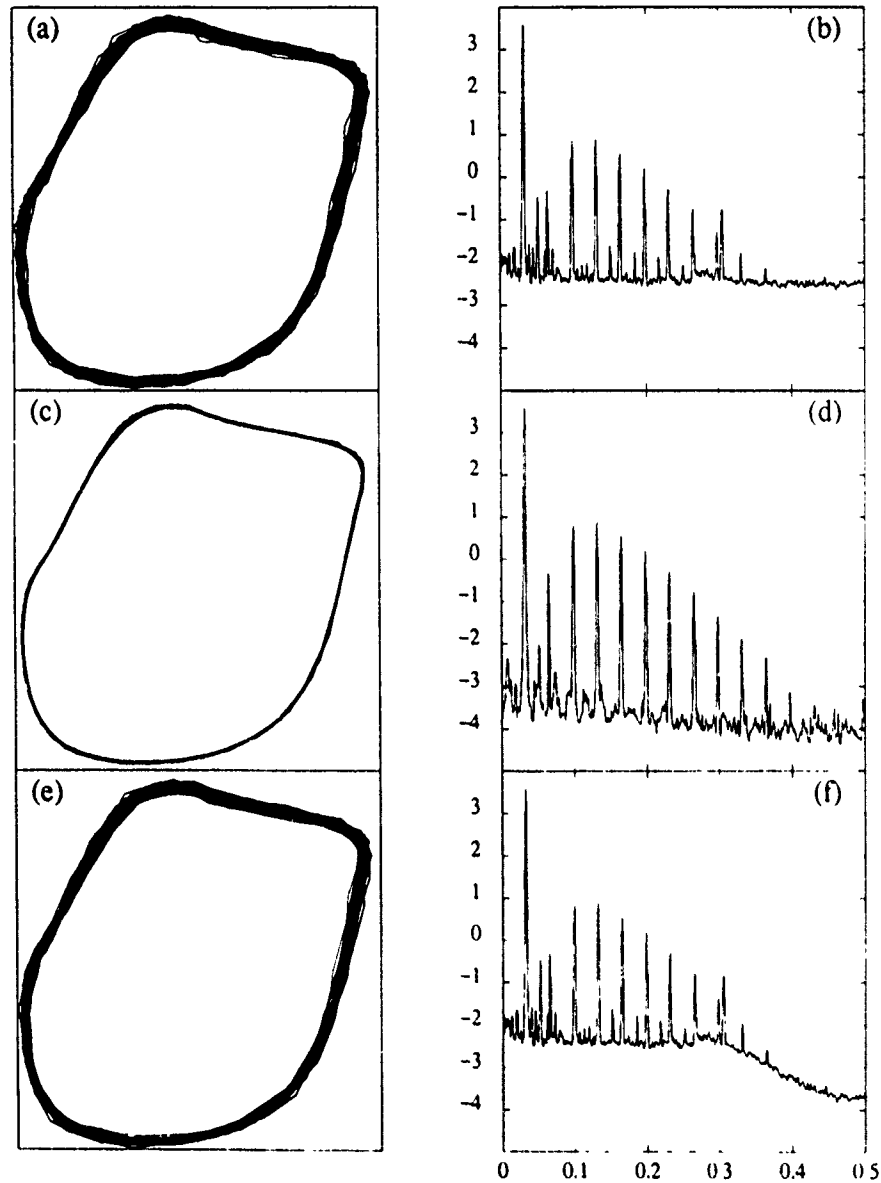


Fig. 4 Phase portraits and power spectra for measurements of wavy vortex flow in a Couette-Taylor experiment. (a), (b) Phase portrait and power spectrum before noise reduction is applied; (c), (d) after noise reduction; (e), (f) after a low-pass filter is applied to the original data. The vertical axis in (b), (d) and (f) is the base-10 logarithm of the power spectral density, the horizontal axis is in multiples of the Nyquist frequency.

These results are significantly different from those obtained by low-pass filtering. Figs. 4e, 4f show the phase portrait and power spectrum when the original data are passed through a 12th-order Butterworth filter with a cutoff frequency of 0.35. The dynamical noise reduction procedure is more

effective than low-pass filtering since the noise appears to have a broad spectrum.

However, the dynamical noise reduction method appears to subtract power from a mode whose fundamental frequency is approximately 0.3 times the Nyquist frequency. We do not know exactly

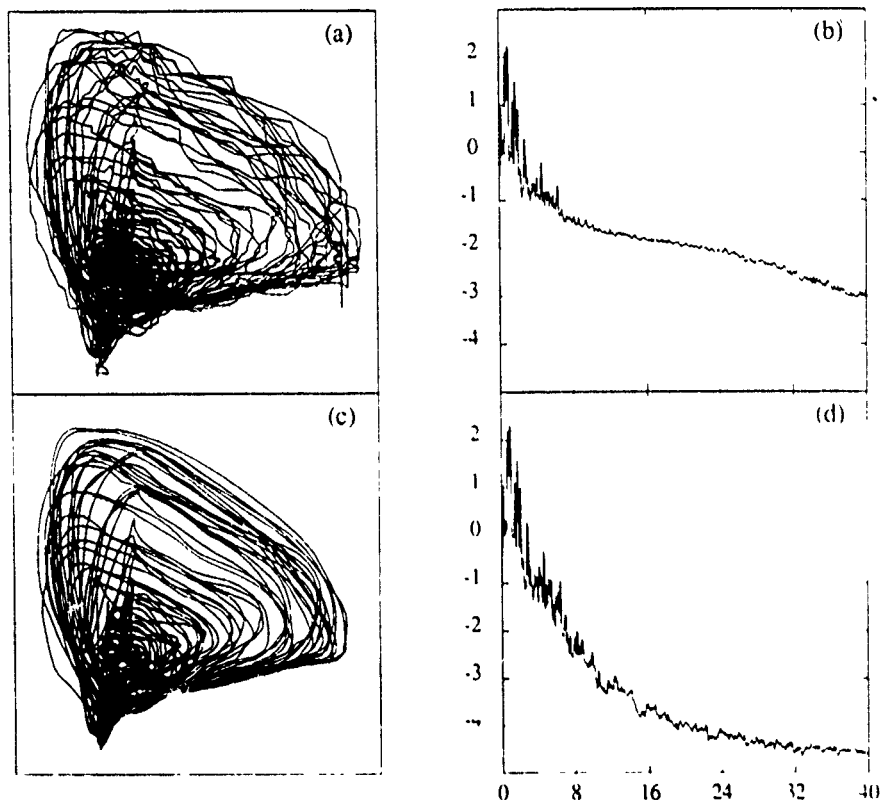


Fig. 5 Phase portraits and power spectra for measurements of weakly chaotic flow in a Couette-Taylor experiment. (a), (b) Phase portrait and power spectrum before noise reduction is applied, (c), (d) after noise reduction. The units for the power spectrum plots are the same as those in ref. [5].

why this occurs. However, this peak corresponds to the rotation frequency of the inner cylinder and may result from a defect in the Couette-Taylor apparatus [33]. We do not consider this to be a serious problem, because the power associated with this mode is several orders of magnitude smaller than that of the wavy vortex flow.

We emphasize that our objective is to find a simple dynamical system that is consistent with the data. It is possible for this method to eliminate certain dynamical behavior from an attractor if those dynamics have very small amplitude, as fig. 4f shows. This situation is most likely to arise when there are not enough data to distinguish such dynamics from random noise. In the present

example, the noise reduction procedure reveals the limit cycle behavior quite well\*\*.

The results obtained by applying the method to chaotic data from the Couette-Taylor fluid flow experiment described in ref. [5] are shown in fig. 5. Fig. 5a shows a two-dimensional phase portrait of the raw time series at a Reynolds number  $R/R_c = 12.9$ , which corresponds to weakly chaotic flow [5]. The corresponding phase portrait from the filtered time series is shown in fig. 5b. Figs. 5c, 5d show

\*\*We have not attempted to find the smallest amplitude at which the noise reduction procedure can distinguish quasiperiodic from periodic flow. In general this will depend on the amount of data, the sampling rate, the embedding dimension, and other factors.

the power spectra for the corresponding time series<sup>\*9</sup>.

It is difficult to estimate how much noise is removed from the data in this example on the basis of power spectra. One problem is that the transition from quasiperiodic to weakly chaotic fluid flow is marked by a sudden rise in the noise floor in the power spectrum (cf. fig. 3 in ref. [5]). Hence one cannot determine how much of the noise floor is due to deterministic chaos and how much results from broad-band noise. The noise reduction procedure described here has the effect of reducing the power in the high-frequency components of the signal. One question therefore is whether reducing the high-frequency noise corresponds to discovering the true dynamics which have been masked by noise. We believe that the answer is yes, based on those cases where there is an underlying low-dimensional dynamical system. However, in chaotic processes some high-frequency components remain, because they are appropriate to the dynamics.

## 7. Numerical experiments on noise reduction

One important question is how much noise this method removes from the data. The power spectra above suggest that the method eliminates most of the noise, but it is impossible to give a precise estimate for typical chaotic experimental data.

However, the Hénon map [19] provides a convenient way to quantify the noise reduction, because it can be written as a time delay map of the form

$$x_{i+1} = f(x_i, x_{i-1}) = 1 - \alpha x_i^2 + \beta x_{i-1}. \quad (3)$$

We use eq. (3) to generate a time series as follows (with the standard parameter values  $\alpha = 1.4$ ,  $\beta = 0.3$ ). We choose an initial condition and discard the first 100 iterates. The next 32768 iterates are

<sup>\*9</sup>The time series consists of 32768 values, from which an attractor is reconstructed in four dimensions. Linear maps are computed using 50–100 points in each ball. Trajectories are fitted using sequences of 24 points.

stored, and a time series is generated by adding a uniformly distributed random number to each iterate. This simulates a time series with *measurement noise*, i.e., a time series where noise results from errors in measuring the signal, not from perturbations of the dynamics.

We measure the improvement in the signal after processing by considering the *pointwise error*

$$e_i = \|x_{i+1} - f(x_i, x_{i-1})\|,$$

i.e., the distance between the observed image and the predicted one. Let the *mean error* be

$$E = \left( \frac{\sum e_i^2}{N} \right)^{1/2},$$

the rms value of the pointwise error over all  $N$  points on the attractor. We define the *noise reduction* as

$$R = 1 - E_{\text{fitted}}/E_{\text{noisy}},$$

where the mean errors are computed for the adjusted and original noisy time series, respectively. The quantity  $R$  is a measure of the self-consistency of the time series. (In other words,  $R$  measures how much better on the average the output attractor obeys eq. (3) as one hops from point to point.)

When 1% noise is added to the input as described above, the noise reduction (measured with the actual map) is 79%<sup>\*10</sup>. Nearly identical results are obtained when the input contains only 0.1% noise. In addition, noise levels can be reduced almost as much in cases where the noise is added to the dynamics, i.e., where the input is of the form  $\{x_{i+1} : x_{i+1} = f(x_i + \eta_i, x_{i-1} + \eta_{i-1}), \eta_i, \eta_{i-1} \text{ random}\}$ . When the program is run on noiseless input, the mean error in the output is 0.025% of the attractor extent, which suggests that errors

<sup>\*10</sup>The pointwise error is measured using eq. (3). However, the attractor can be embedded in more than two dimensions when performing the noise reduction.

arising from small nonlinearities are negligible when the input contains enough points.

## 8. Simplicial approximations of dynamical systems

Recent work has shown that simplicial approximations of dynamical systems can reproduce the behavior of the original system to high accuracy [36]. (See also ref. [35] for a bilinear approach.) In particular, the fractal structure of the original attractors and basin boundaries is preserved over many scales. Such approximations can yield significant computational savings, especially when the original system consists of ordinary differential equations.

This approach can be extended in a natural way to generate simplicial approximations of the dynamics on attractors reconstructed from experimental data. Our objective here is to find an approximate dynamical system in a neighborhood of the attractor as follows.

A *simplex* in an  $m$ -dimensional space is a triangle with  $m + 1$  vertices. Suppose the map is known at each point on a grid. Then there is a unique way to extend the map linearly to the interior of the simplex  $S$  whose vertices are grid points. Given a point  $P$  in the interior of  $S$ , let  $\{b_i\}_{i=0}^m$  be its corresponding *barycentric coordinates* (see ref. [36] for an algorithm to compute them). Let  $f(v_i)$  be the map at the  $i$ th vertex. The dynamical system at  $P$  is approximated by computing

$$\Phi(P) = \sum_{i=0}^m b_i f(v_i). \quad (4)$$

We apply this method to experimental data by finding a linear approximation of the dynamics at each vertex  $v_i$  with the least-squares method described above, using a collection of points in a small ball around  $v_i$ . The maps are stored and retrieved using a hashing algorithm similar to that described in ref. [36]. This yields a piecewise linear approximation of the dynamics from a set of experimental data which can be analyzed with the

methods that previously were available only to theorists<sup>#11</sup>.

We illustrate the approach using a time series of 32 768 values from the Hénon map with  $\alpha = 1.2$ ,  $\beta = 0.3$  using eq. (3) and adding 0.1% noise as described above. The original attractor is shown in fig. 6a. We take a grid of points which are spaced at 1% intervals (this and subsequent distances are expressed as a fraction of the original attractor extent). The time series is embedded in two dimensions, and a linear approximation of the dynamics is computed at each grid point for which 50 or more attractor points can be collected with a ball of radius 0.03; the set of such grid points is shown in fig. 6b. We take an initial condition near the original attractor and show the first 3000 iterates using eq. (4) in fig. 6c. Although some defects are visible, the attractor produced by the approximate dynamical system looks almost identical to the original one.

One application of simplicial approximations is the location of periodic saddles and the estimation of the largest eigenvalue of the corresponding Jacobian. That is, if  $x$  is a periodic point of period  $p$ , then we find the eigenvalue of  $Df^p(x)$  of largest modulus, where  $Df^p(\cdot)$  refers to the matrix of partial derivatives of the  $p$ th iterate of the map  $f$  evaluated at  $x$ .

Given an initial guess for  $x$ , one can apply Newton's method using the maps computed at the grid points and eq. (4) to locate the saddle using the simplicial approximations. Likewise, eq. (3) can be used to locate the corresponding "exact" saddle. Saddle orbits up to period 8 have been computed in this way. In all cases, the saddle point for the simplicial approximation is within 2% of the corresponding saddle point for the Hénon map. Table 1 shows the largest eigenvalues of the saddle orbits. (The columns labeled  $m = 2$  and  $m = 3$  refer to the embedding dimension used to reconstruct the attractor.) In most cases, the

<sup>#11</sup>This approach is less ambitious than that of Crutchfield [8], who attempts to find a single set of nonlinear difference equations that creates the observed attractor

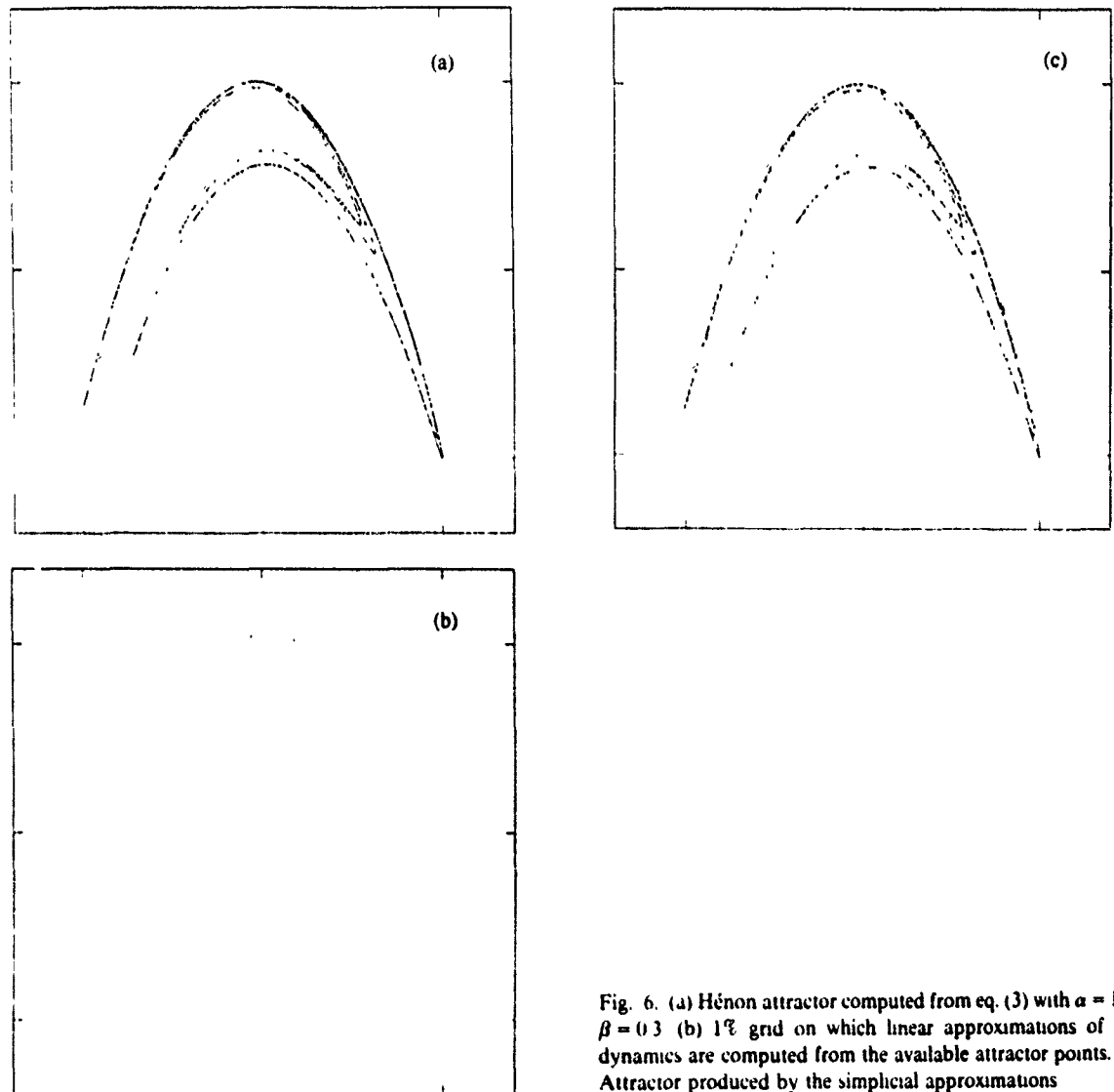


Fig. 6. (a) Hénon attractor computed from eq. (3) with  $\alpha = 1.2$ ,  $\beta = 0.3$ . (b) 1% grid on which linear approximations of the dynamics are computed from the available attractor points. (c) Attractor produced by the simplicial approximations

relative error is only a few percent, and in no case exceeds 25%. (The largest relative error is for the period 8 saddles, where one finds the eigenvalue of the product of 8 Jacobians computed from the least squares.)

This method can be extended to experimental data sets. However, there are relatively stringent requirements on the data that can be handled: the time series must be long enough to trace out many trajectories near the principal unstable saddle orbits, and the noise level must be low. (Presumably, noisy data can be preprocessed using the approach

described in section 4.) The current computer implementation uses a large amount of disk space to store the linear map approximations at the grid points.

We have constructed a simplicial approximation for an attractor obtained from a Belousov-Zhabotinskii chemical reaction [7, 30]. The attractor is reconstructed in three dimensions from a set of 32 768 measurements of bromide ion concentration. The phase portrait is shown in fig. 7a.

Linear approximations of the dynamics are computed at each point of a grid consisting of 50

Table 1  
The largest eigenvalues of the Jacobian of the periodic orbits located using the simplicial approximation of the Hénon attractor.

Period	$m = 2$	Exact	$m = 3$
1	1.793	1.695	1.757
2	2.178	2.199	2.183
4	4.226	4.329	4.051
6	10.38	10.70	9.626
6	10.38	11.32	12.12
8	25.80	24.88	30.25
8	20.02	20.60	20.38
8	17.70	24.32	21.70

intervals along each coordinate axis for which 50 or more attractor points can be located within an 8% radius of the grid point. This produces a database of 59 550 maps. We observe from graphical evidence that many trajectories approach what appears to be a period-3 saddle in the middle of

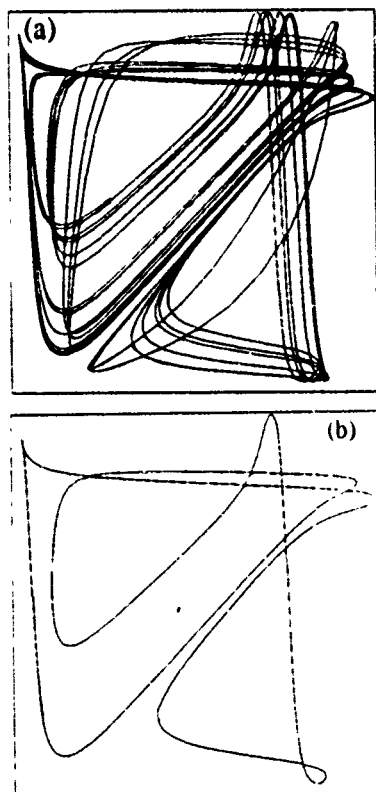


Fig. 7. (a) The attractor reconstructed from a time series of bromide ion concentrations in a Belousov-Zhabotinsky chemical reaction. (b) The period-3 saddle orbit

the attractor. Using initial guesses from some of the trajectories, we apply Newton's method to locate the saddle orbit shown in fig. 7b. Moreover, we obtain estimates of the Jacobian  $Df$  of the map evaluated at a point on the saddle orbit. The eigenvalues of  $Df$  are estimated as  $\lambda_1 = 1.14$ ,  $\lambda_2 = 0.102$ , and  $\lambda_3 = -1.53$ . These quantitative results confirm that the orbit is a saddle since  $\lambda_1 > 0 > \lambda_3$ . (Note that one expects  $\lambda_2 = 0$  for a flow generated from a set of differential equations.)

## 9. Conclusion

Methods for approximating the dynamics of attractors reconstructed from experimental data provide powerful tools. Most of the same procedures that have been so important for theoretical insight, such as Poincaré maps, unstable fixed points and their manifolds, basin boundaries, and the like, are now available to experimenters, at least in cases where the dynamics are low dimensional. There is little doubt that these tools will lead to breakthroughs in the understanding of a wide variety of physical systems. However, considerable effort is needed before we learn which kinds of systems will benefit most from these types of analyses. Significant improvements in technique will certainly extend the applicability of dynamical embedding methods, for example to higher-dimensional attractors.

## Appendix

In this appendix we outline a possible alternative noise reduction method based on the theory of least squares when all the quantities in the regression are measured with error.

In ordinary least squares, the variables in the problem fall into two classes: the *independent* variables, which are known exactly, and the *dependent* variables, which are observations assumed to be functions of the independent variables. The dependent variables are subject to random errors that are assumed independent and identically distributed (i.i.d.).

On an attractor reconstructed from experimental data, we assume that the mapping which takes points in a sufficiently small ball to their images is approximately linear. However, the locations of all the points are subject to small random errors because of the noise. Hence one cannot describe the points as independent variables and their images as dependent variables. The usual least-squares method produces a biased estimate of the linear map, and this bias does not decrease if more observations are added [16, 23].

The so-called "errors in variables" least-squares methods can be used to handle the latter problem. This approach can be used to obtain both an estimate of the linear map as well as estimates of the "true" values of each of the observations.

At first this appears to be an underdetermined problem: from  $n$  pairs of observations one wants to compute the parameters of the functional relation between them as well as estimates of the  $n$  actual pairs<sup>\*12</sup>. However, it is possible to solve this problem by making some assumptions about the errors [16, 23].

In our case, we assume that the errors in the location of each point and its image are i.i.d. In particular, we let the covariance matrix of the errors in the variables be the identity matrix. This assumption is valid whenever the noise is independent of the dynamics<sup>\*13</sup>.

We illustrate the procedure for the case where we are given a collection of  $n$  points (in  $\mathbb{R}^m$ ) and their images. Following Jefferys [21], we form a set of  $n$  equations of condition given by

$$f_i(x_i) = x_{n+i} - Ax_i - b_i \equiv x_{n+i} - L(x_i), \quad (5)$$

where  $x_i$  is the  $i$ th point,  $x_{n+i}$  is its observed image,  $A$  is an  $m \times m$  matrix, and  $b$  is an  $m$ -vector. The goal is to find estimates of  $L$  (i.e.,  $A$  and

$b$ ), together with perturbations  $\hat{v}$ , such that

$$f_i(x_i + \hat{v}_i) = (x_{n+i} + \hat{v}_{n+i}) - L(x_i + \hat{v}_i) \approx 0$$

and such that the quadratic form

$$S_0 = \frac{1}{2} \hat{v}^t \sigma^{-1} \hat{v} \quad (6)$$

is minimized. The superscript  $t$  denotes transpose and  $\sigma$  is the covariance matrix of the observations (which we assume is the identity matrix here).

This minimization problem can be solved using Lagrange multipliers (see refs. [21, 22] for a numerical algorithm). The solution gives  $A$  and  $b$  together with estimates  $x_i + \hat{v}_i$  of the "true" observations. It can be shown [16] under fairly mild hypotheses that the estimates of  $L$  and the observations are the best in the class of linear estimators.

One way to approach noise reduction is to extend eq. (5) to include several iterations of the observed points. Given a collection of points in a ball, together with the next  $p$  iterates of each point, the method above is used to find a collection of linear maps  $L_1, L_2, \dots, L_p$  approximating the dynamics. The method also finds estimates of the actual observations. In this approach, therefore, the calculation of the maps and the adjustment of the trajectories is done in one step. Moreover, each point and its image exactly satisfy a linear relationship.

Of course,  $p$  cannot be too large, because nonlinear effects eventually will become significant when the dynamics are chaotic. On the other hand, eq. (5) provides a natural way to include quadratic or other nonlinear terms.

We have written a computer program to implement this alternative noise reduction algorithm. So far, the results of this approach have not been as good as those from the method described in the main part of the paper, but further refinement should improve them.

## Acknowledgements

Dan Lathrop provided invaluable assistance in finding periodic orbits in the Hénon and BZ attractors. We thank Bill Jefferys for useful discus-

<sup>\*12</sup>In the statistical literature, the problem is said to be *unidentified*.

<sup>\*13</sup>Dynamical noise (i.e., each point is perturbed slightly before iterating) yields a covariance matrix which depends on the point. However, as long as the dynamical noise is small, our assumptions about the covariance matrix of the errors should not compromise the accuracy of the method.

sions and computer software for the errors in variables least-squares problem. Andy Fraser, Randy Tagg and Harry Swinney all provided helpful suggestions. This research is supported by the Applied and Computational Mathematics Program of the Defense Advanced Research Projects Agency (DARPA-ACMP) and by the Department of Energy Office of Basic Energy Sciences.

## References

- [1] R. Badu, G. Brogg, D. Denghetti, M. Ravan, S. Ciliberto and A. Politi, *Phys. Rev. Lett.* 60 (1988) 979
- [2] R.P. Behringer and G. Ahlers, *J. Fluid Mech.* 125 (1982) 219;  
G. Ahlers and R.P. Behringer, *Phys. Rev. Lett.* 40 (1978) 712.
- [3] J.L. Bentley and J.H. Friedman, *ACM Comput. Surv.* 11 (1979) 397.
- [4] S.M. Bozic, *Digital and Kalman Filtering* (Edward Arnold Publishers Ltd., London, 1979).
- [5] A. Brandstätter and H.L. Swinney, *Phys. Rev. A* 35 (1987) 2207.
- [6] M. Casdagli, *Nonlinear prediction of chaotic time series*, preprint (December 1987).
- [7] K.C. Coffman, Ph.D. Thesis, University of Texas at Austin (1987)
- [8] J.P. Crutchfield and B. McNamara, *Complex Systems* 1 (1987) 417;  
H.D. Abarbanel, R. Brown and J.B. Kadtko, *Prediction and system identification in chaotic nonlinear systems: time series with broadband spectra*, preprint (January 1989).
- [9] J.J. Dongarra, C.B. Moler, J.R. Bunch and G.W. Stewart, *LIN-PACK User's Guide* (Society for Industrial and Applied Mathematics, Philadelphia, 1979).
- [10] J.-P. Eckmann, S.O. Kamphorst, D. Ruelle and S. Ciliberto, *Phys. Rev. A* 34 (1986) 4971
- [11] J.-P. Eckmann and D. Ruelle, *Rev. Mod. Phys.* 57 (1985) 617.
- [12] J.D. Farmer and J.J. Sidorowich, *Phys. Rev. Lett.* 59 (1987) 845
- [13] P.R. Fenstermacher, H.L. Swinney and J.P. Gollub, *J. Fluid Mech.* 94 (1979) 103.
- [14] W. Franceschini and L. Russo, *J. Stat. Phys.* 25 (1981) 757.
- [15] A. Fraser and H.L. Swinney, *Phys. Rev. A* 34 (1986) 1134.
- [16] W.A. Fuller, *Measurement Error Models* (Wiley, New York, 1987).
- [17] E.G. Gwinn and R.M. Westervelt, *Phys. Rev. A* 33 (1986) 4143.
- [18] S.M. Hammel, J.A. Yorke and C. Grebogi, *J. Complexity* 3 (1987) 136; *Bull. Am. Math. Soc.* 19 (1988) 465
- [19] M. Hénon, *Comm. Math. Phys.* 50 (1976) 69.
- [20] D. Hirst, Ph.D. Dissertation, University of Texas (December 1987).
- [21] W.H. Jefferys, *Astron. J.* 85 (1980) 177
- [22] W.H. Jefferys, *Astron. J.* 86 (1981) 149
- [23] M.G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 2 (Griffin, London, 1961) p. 375
- [24] A. Libchaber, S. Fauve and C. Laroche, *Physica D* 7 (1983) 73.
- [25] E.N. Lorenz, *J. Atmos. Sci.* 20 (1963) 130
- [26] S.W. MacDonald, C. Grebogi, E. Ott and J.A. Yorke, *Physica D* 17 (1985) 125.
- [27] G. Mayer-Kress, ed., *Dimensions and Entropies in Chaotic Systems* (Springer, Berlin, 1986), and references therein.
- [28] F. Mitschke, M. Möller and W. Lange, *Phys. Rev. A* 37 (1988) 4518.
- [29] L.R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1975).
- [30] J.-C. Roux, *Physica D* 7 (1983) 57;  
J.-C. Roux, R.H. Simoyi and H.L. Swinney, *Physica D* 8 (1983) 257.
- [31] M. Sano and Y. Sawada, *Phys. Rev. Lett.* 55 (1985) 1082.
- [32] C. Sparrow, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors* (Springer, New York, 1982).
- [33] R. Tagg, private communication.
- [34] F. Takens, in: *Dynamical Systems and Turbulence*, Springer Lecture Notes in Mathematics, Vol. 898, eds. D.A. Rand and L.-S. Young (Springer, Berlin, 1980) p. 366
- [35] B.H. Tongue, *Physica D* 28 (1987) 401
- [36] F. Varosi, C. Grebogi and J.A. Yorke, *Phys. Lett. A* 124 (1987) 59
- [37] A. Wolf, J.B. Swift, H.L. Swinney and J.A. Vastano, *Physica D* 16 (1985) 285.

## ANTIMONOTONICITY: CONCURRENT CREATION AND ANNIHILATION OF PERIODIC ORBITS

I. KAN AND J. A. YORKE

**ABSTRACT.** One-parameter families  $f_\lambda$  of diffeomorphisms of the Euclidean plane are known to have a complicated bifurcation pattern as  $\lambda$  varies near certain values, namely where homoclinic tangencies are created. We argue that the bifurcation pattern is much more irregular than previously reported. Our results contrast with the monotonicity result for the well-understood one-dimensional family  $g_\lambda(x) = \lambda x(1-x)$ , where it is known that periodic orbits are created and never annihilated as  $\lambda$  increases. We show that this monotonicity in the creation of periodic orbits never occurs for any one-parameter family of  $C^3$  area contracting diffeomorphisms of the Euclidean plane, excluding certain technical degenerate cases where our analysis breaks down. It has been shown that in each neighborhood of a parameter value at which a homoclinic tangency occurs, there are either infinitely many parameter values at which periodic orbits are created or infinitely many at which periodic orbits are annihilated. We show that there are *both* infinitely many values at which periodic orbits are *created* and infinitely many at which periodic orbits are *annihilated*. We call this phenomenon *antimonotonicity*.

### I. INTRODUCTION

The orbit of point  $x$  under a diffeomorphism of the plane  $f$  is the sequence  $\{f^k(x)\}$ , where for  $k \geq 0$ ,  $f^k$  denotes the  $k$ -fold composition of  $f$ ,  $f^{-k}$  denotes the  $k$ -fold composition of  $f^{-1}$  and  $f^0$  is the identity map. Let  $p$  be a periodic point with period  $n$ . The stable manifold  $W^s(p)$  of the point  $p$  is the set  $\{x : \lim_{k \rightarrow \infty} f^{nk}(x) = p\}$ . Similarly, the unstable manifold  $W^u(p)$  of  $p$  is  $\{x : \lim_{k \rightarrow \infty} f^{-nk}(x) = p\}$ . We assume that  $p$  is a hyperbolic saddle, that is, the eigenvalues  $e_1, e_2$  of  $Df^n(p)$  are such that  $|e_1| < 1 < |e_2|$ . Since  $f$  is a diffeomorphism of the plane, both  $W^s(p)$  and  $W^u(p)$  are curves. There exists a homoclinic tangency

Received by the editors November 2, 1988 and, in revised form, May 1, 1989.  
 1980 *Mathematics Subject Classification* (1985 Revision). Primary 54C35, 58F13.  
 Partial support provided by DARPA/ACMP program and by AFOSR-81-0217.

of  $p$  at  $q$  if  $W^s(p)$  and  $W^u(p)$  intersect tangentially at  $q$ . The homoclinic tangency of  $p$  at  $q$  for a one-parameter family  $f_\lambda$  at  $\lambda = \lambda_0$  is called *nondegenerate* if  $W^s(p)$  and  $W^u(p)$  have quadratic contact at  $q$  and  $W^s(p)$  has nonzero velocity transverse to  $W^u(p)$  at  $q$  as  $\lambda$  varies [R]. Any value  $\lambda_0$  at which this occurs is called a *nondegenerate tangency value*.

A one-parameter family of maps  $g_\lambda$  is called *monotone increasing* (*decreasing*) on an interval  $J$  of parameter values if there are no bifurcations for  $\lambda \in J$  in which periodic orbits are annihilated as  $\lambda$  increases (decreases, respectively). We say  $f_\lambda$  is *antimonotone* at  $\lambda_0$  if periodic orbits are both created and annihilated as  $\lambda$  increases in each neighborhood of the parameter value  $\lambda_0$ .

The only smooth family for which monotonicity has been proved is the quadratic family  $g_\lambda(x) = \lambda x(1-x)$  [Douady, Hubbard, Milnor, Thurston, Sullivan, see [MT]]. By contrast we have the following theorem.

**Antimonotonicity Theorem.** *Each dissipative  $C^3$  planar diffeomorphism family is antimonotone at each nondegenerate homoclinic tangency value.*

Note that this result says nothing about what happens near degenerate homoclinic tangency values, but we believe this situation is essentially the same as for the nondegenerate case.

We sketch the proof for a model case. A paper detailing the proof of the general result is in preparation. If two curves are tangent at  $\lambda = \lambda_0$  and move apart, so that they do not intersect as  $\lambda$  increases (decreases) beyond  $\lambda_0$ , then we say *contact is broken* at  $\lambda_0$  (*contact is made* at  $\lambda_0$ , respectively), and we say  $\lambda_0$  is a *contact-breaking value* (*contact-making value*, respectively).

**Bubble Lemma.** *If  $\lambda_0$  is a nondegenerate tangency value at which contact is made, then there are nondegenerate tangency values arbitrarily close to  $\lambda_0$  at which contact is broken (and vice versa).*

The theorem follows immediately from the Bubble Lemma because in each neighborhood of a contact-making nondegenerate tangency value, infinitely many periodic orbits are created (and near contact-breaking ones, infinitely many are annihilated) [N, GS]. Thus, in each neighborhood of a nondegenerate tangency, orbits are both created and annihilated, as is illustrated in Figure 1 for the example of the Henon family.



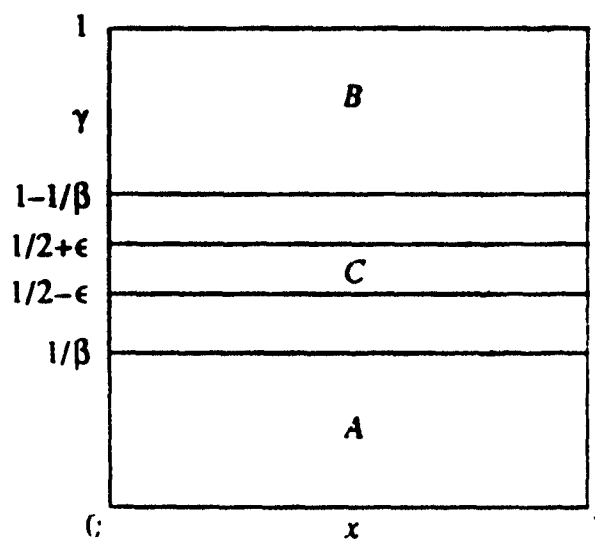
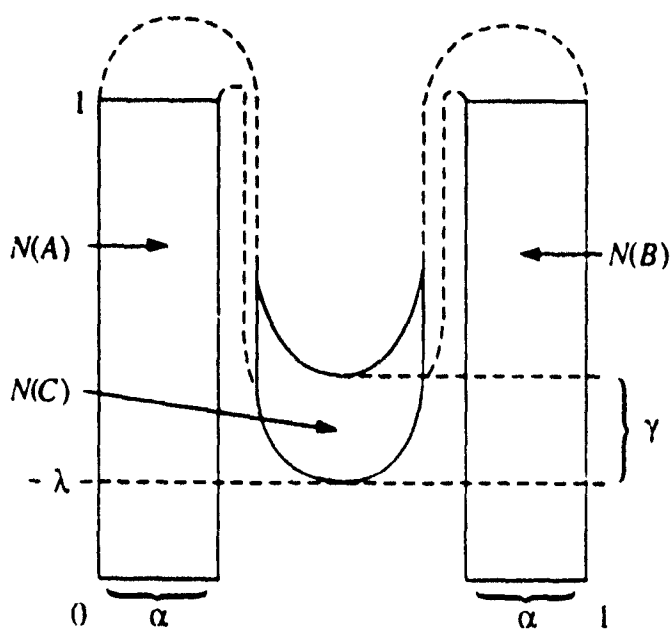
FIGURE 1. SMALL BUBBLE IN HENON FAMILY  $H_\lambda(x, y) = \lambda - y^2 + 0.3y, x)$ . 5,000 PREITERATES. A-COORDINATE OF 80,000 ITERATES PLOTTED PER  $\lambda$  VALUE.

## II. PRELIMINARIES

For each Cantor set  $C \subset \mathbb{R}$  Newhouse [N] defines a number in  $[0, \infty)$  called the thickness  $\tau(C)$  associated with  $C$ . A "middle- $\theta$ " Cantor set  $C_\theta = I \setminus G_\theta$  is constructed inductively as follows:  $I = [0, 1]$  and  $I_{1,0}$  and  $I_{1,1}$  are the left and right component of  $I \setminus G_\theta$ , respectively, where  $G_\theta$  is an open interval of length  $\theta \cdot |I_1|$  in the middle of  $I_1$ . The thickness of  $C_\theta$  is  $(1 - \theta)/2\theta$ . Newhouse proves the following lemma.

**Thickness Lemma.** *Let  $F$  and  $H$  be Cantor sets in  $\mathbb{R}$ , with  $H \cap \text{hull}(F)$  and  $\text{hull}(H) \cap F$  both nonempty, and  $\tau(H) \cdot \tau(F) > 1$ . Then  $H \cap F$  is nonempty.*

A Newhouse horseshoe family  $N_\lambda$  is defined as follows. (See Figure 2 on page 472 for symbols, coordinates, and the role of the constants, and see Figure 3 on page 472 for the first iterate

FIGURE 2. COORDINATES FOR  $N_{\lambda}$ .FIGURE 3. FIRST ITERATE OF  $N_{\lambda}$ .

$N_{\lambda}$ . Define  $N_{\lambda}(x, y) = (\alpha x, \beta y)$  for  $(x, y) \in A$ ;  $N_{\lambda}(x, y) = (1 - \alpha x, \beta(1 - y))$  for  $(x, y) \in B$ ;  $N_{\lambda}(x, y) = (y, -\lambda + \gamma(1 - x) + \delta(y - 1/2)^2)$  for  $(x, y) \in C$ ; and continue  $N_{\lambda}$  smoothly to the rest of  $\mathbb{R}^2$ .

We choose  $\alpha\beta < 1$  so  $N_{\lambda}$  is dissipative (i.e.  $|\det D(N_{\lambda})| < 1$ ) throughout  $A \cup B$ , and we choose  $\alpha, \beta, \gamma, \delta, \varepsilon$  such that  $N_{\lambda}$  is one-to-one on  $A \cup B \cup C$ . This implies  $\beta > 2$ . Let  $\Lambda$  denote the maximal invariant subset of  $A \cup B$ ;  $\Lambda$  is a Cantor set and is the product  $\Lambda_u \times \Lambda_s$  of two Cantor sets.  $\Lambda_u$  is the projection of  $\Lambda$

onto the  $x$ -axis and  $\Lambda_s$  onto the  $y$ -axis. We assume that  $\alpha$  and  $\beta$  are selected so that  $\tau(\Lambda_s) \cdot \tau(\Lambda_u) = (\beta - 2)^{-1} \alpha (1 - 2\alpha)^{-1} > 1$ .

A *primary stable (unstable) segment* is a line segment of the form  $[0, 1] \times \{y\}$  where  $y \in \Lambda_s$ , ( $\{x\} \times [0, 1]$  where  $x \in \Lambda_u$ , respectively). A *primary unstable parabola* is a parabolic arc of the form  $N_\lambda(x, [1/2 - \varepsilon, 1/2 + \varepsilon])$  where  $x \in \Lambda_u$ .

Newhouse and Robinson show in [N, R], that in effect, there exist parameter values  $\lambda$  near homoclinic tangencies where for a proper choice of coordinates the map is similar to Figure 3. We are assuming that the map changes in a regular way as  $\lambda$  varies, thereby avoiding technical complications.

### III. PROOF OF BUBBLE LEMMA

#### ASSUMING NEWHOUSE HORSESHOE FAMILIES OCCUR

Let  $\lambda_0$  be a nondegenerate tangency value, which we assume to be a contact-making tangency. We assume that on a small interval, arbitrarily near  $\lambda_0$ , there is a Newhouse horseshoe family. We rescale that small interval to be  $[0, 1]$ . The primary tangencies (the tangencies of primary parabolas with primary stable segments) are all contact making. We will show that arbitrarily near  $\lambda = 0$ , there is a nondegenerate tangency which is contact-breaking and is not primary.

The parabolic arc of the form  $v(t, \xi) = (1/2 + t, \beta^{-n} + \xi + \delta t^2)$  for  $0 < \xi < (1 - 2/\beta)\beta^{1-n}$ ,  $\delta t^2 < \beta^{-n} - 2\beta^{-n-1} - \xi$ ,  $|t| < \varepsilon$ , lies in a gap in the Cantor set of primary stable leaves as shown in Figure 4.

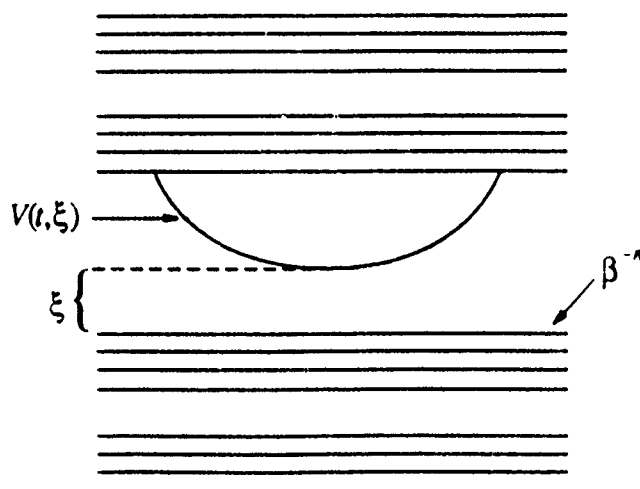


FIGURE 4. THE ARC  $v(t, \xi)$ .

Let  $\Gamma(\xi)$  denote the set of parameters such that  $v(t, \xi)$  lies on a primary parabola. For each  $\lambda$  the vertices of the primary parabolas have  $y$ -coordinates at  $(-\lambda + \gamma\Lambda_u)$ , so we see that  $\Gamma(\xi) = -\gamma\Lambda_u - \xi - \beta^{-n}$  and the thickness of  $\Gamma(\xi)$  is equal to  $\tau(\Lambda_u)$ . The  $n$ th image of  $v(t, \xi)$  under  $N_\lambda$  is

$$\begin{aligned} v_n(t, \xi, \lambda) = & (\beta^{-1} - \xi\beta^{n-1} - \delta\beta^{n-1}t^2 \\ & - \lambda - \gamma\alpha^{n-1}(1/2 + t) + \gamma \\ & + \delta(\beta^{-1} - \xi\beta^{n-1} - \delta\beta^{n-1}t^2 - 1/2)^2). \end{aligned}$$

There is a  $\xi = \bar{\xi}$ ,  $t = \bar{t}$  at which the  $y$ -coordinate has a stationary inflection point as shown in Figure 5b, and  $\bar{\xi}$  and  $\bar{t}$  satisfy  $4\delta\beta^{n-1}(\beta^{-1} - \bar{\xi}\beta^{n-1} - 1/2) = -3(\gamma\alpha^{n-1}\beta^{n-1/2})^3$ , and  $\bar{t} = -\gamma\alpha^{n-1}\beta^{n-1/2}/4\delta\beta^{n-1}$ . Notice

$$\bar{\xi} = (1/2)(1 - 2/\beta)\beta^{1-n} + (\alpha\beta)^{2n/3}O(\beta^{-2n}),$$

so for large  $n$  we have  $0 < \bar{\xi} < (1 - 2/\beta)\beta^{1-n}$ .

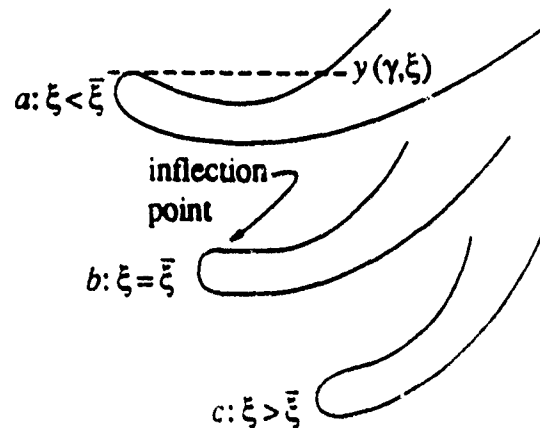
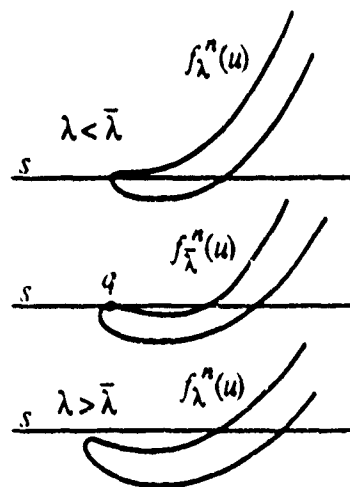
*Claim.* For fixed  $\xi \leq \bar{\xi}$ , with  $\bar{\xi} - \xi$  sufficiently small, there exists a  $\lambda \in \Gamma(\xi)$  such that the  $n$ th iterate of the primary parabola containing  $v(t, \xi)$  has a tangency with a primary stable segment. This tangency is contact-breaking and is nondegenerate for  $\xi < \bar{\xi}$ .

The first part of this claim follows from the fact that the local maximum  $v(\xi, \lambda)$  (see Figure 5a) of the  $y$ -coordinate of  $v_n(t, \xi, \lambda)$  depends linearly on  $\lambda$ . That is,

$$\{v(\xi, \lambda) | \lambda \in \Gamma(\xi)\} = \{v(\xi, 0) - \lambda | \lambda \in \Gamma(\xi)\},$$

and so  $\{v(\xi, \lambda) | \lambda \in \Gamma(\xi)\}$  has thickness  $\tau(\Lambda_u)$ . By the Thickness Lemma, there exists some  $\hat{\lambda} \in \Gamma(\xi)$  such that  $v(\xi, \hat{\lambda}) \in \Lambda_u$ . Note that  $\hat{\lambda}$  is  $O(\beta^{-n})$ . Since  $\hat{\lambda}$  is in  $\Gamma(\xi)$ , there is a primary unstable parabola which contains  $v(t, \xi)$ , so  $v_n(t, \xi, \hat{\lambda})$  is contained in the unstable manifold of  $\Lambda$  and is tangent to a primary stable segment of  $\Lambda$ . As  $\lambda$  varies near 0, the position of this primary unstable parabola is  $v(t, \xi + \lambda)$ . Nondegeneracy and contact-breaking can be verified by considering the  $y$ -coordinate of  $d(v_n(t, \xi + \lambda, \hat{\lambda}))/d\lambda$  and noting that for sufficiently small  $\bar{\xi} - \xi > 0$  and large  $n$  this derivative is negative for  $t$  sufficiently close to  $\bar{t}$ .

We have shown that there is a primary stable leaf  $S$  and a primary unstable parabola  $U$  so that the  $n$ th iterate of  $U$  has a


 FIGURE 5. THE INFLECTION VALUE  $\bar{\xi}$ .

 FIGURE 6. CONTACT-BREAKING TANGENCY  $q$  AT  $\lambda = \bar{\lambda}$ .

contact-breaking tangency with  $S$  (see Figure 6). Since the stable and unstable manifold of the fixed point  $p$  at  $(0, 0)$  contain curves arbitrarily close to  $S$  and  $U$ , respectively, we see that  $p$  will have contact-breaking tangencies at parameter values arbitrarily near  $\bar{\lambda}$ . Finally, for  $n$  large, this  $\bar{\lambda}$  is near 0.

# REFERENCES

- [GS] N. K. Gavrilov and L. P. Šilnikov, *On three-dimensional dynamical systems close to systems with structurally unstable homoclinic curve*, I, II, Math. USSR-Sb. **88**(4) (1972), 467–485, *ibid.* **90**(1) (1973), 139–156.
- [MT] J. Milnor and W. P. Thurston, *On iterated maps of the interval*, Dynamical Systems: Proc. Univ. Maryland 1986–87, Lecture Notes in Math., vol. 1342, Springer-Verlag, Berlin and New York, 1989, pp. 465–563.

- [N] S. Newhouse. *The abundance of wild hyperbolic sets and nonsmooth stable sets for diffeomorphisms*. Inst. Hautes Etudes Sci. Publ. Math. **50** (1978), 101-151.
- [R] C. Robinson. *Bifurcation to infinitely many sinks*. Comm. Math. Phys. **90** (1983), 433-459.

DEPARTMENT OF MATHEMATICS, GEORGE MASON UNIVERSITY, FAIRFAX, VIRGINIA 22030 AND NAVAL SURFACE WARFARE CENTER, WHITE OAK, MARYLAND

INSTITUTE FOR PHYSICAL SCIENCE AND TECHNOLOGY, UNIVERSITY OF MARYLAND, COLLEGE PARK, MARYLAND 20742

## CHAOTIC SCATTERING IN SEVERAL DIMENSIONS

Qi CHEN, Mingzhou DING<sup>1</sup> and Edward OTT<sup>2</sup>

*Laboratory for Plasma Research, University of Maryland, College Park, MD 20742, USA*

Received 20 December 1989; revised manuscript received 30 January 1990; accepted for publication 30 January 1990

Communicated by A.P. Fordy

For chaotic scattering in two-degree-of-freedom ( $N=2$ ), time-independent, Hamiltonian systems, scattering functions (i.e., plots of the dependence of a phase space variable after scattering versus a phase space variable before scattering) typically display singularities on a fractal set. For  $N>2$ , however, scattering functions typically do not have fractal properties (even when the chaotic invariant set is fractal), unless the fractal dimension of the chaotic set is large enough. A numerical investigation of this phenomenon is presented for a scatterer consisting of four reflecting spheres at the vertices of a regular tetrahedron.

Recently, there has been much interest in the phenomenon of chaotic scattering (see reviews [1]) due to its appearance in a variety of applications, including fluid mechanics, celestial mechanics, and, especially, molecular dynamics. In addition, the implications of classical chaotic scattering for the corresponding quantum scattering problem is a subject of active research [2]. Another line of study concerns the question of how chaotic scattering comes about and evolves as a system parameter is varied [3]. In all of these past works, when specific systems or examples are investigated, they have almost always been effectively Hamiltonians with two degrees of freedom. Since many situations that will arise in practice can be expected to involve Hamiltonians with more than two degrees of freedoms, it is important to see whether new phenomena, not present in two-degree-of-freedom systems, can be anticipated in these situations.

In particular, let us consider plotting variables characterizing the state of the system after scattering as a function of a single variable characterizing the state of the system before scattering (with the other "before-scattering variables" held fixed). We call such plots "scattering functions". It is a striking hallmark of chaotic scattering in two-degree-of-freedom

systems that these functions are typically singular on a Cantor set of values of the variable characterizing the state before scattering. Here we consider whether this situation persists in systems with more than two degrees of freedom. We find that the scattering function does not typically display fractal properties in  $N$ -degree-of-freedom chaotic scattering systems with  $N>2$ , unless the Hausdorff dimension  $D_c$  of the fractal chaotic invariant set exceeds a critical value. In particular, if the Hamiltonian is time reversible, then fractal behavior of scattering functions can typically be expected only if

$$D_c > 2N - 3. \quad (1)$$

Since  $D_c$  is greater than or equal to one, eq. (1) is satisfied for two-degree-of-freedom chaotic scattering systems ( $N=2$ ). For  $N>2$ , fractal behavior of the scattering function is typically always absent even though the chaotic invariant set itself is fractal, provided that  $1 \leq D_c < 2N - 3$ . (Because the chaotic set lies in the  $D_E$ -dimensional energy surface ( $D_E=2N-1$ ), we always have  $D_c \leq 2N-1$ .) Since  $D_c$  depends on system parameters, one expects that a qualitative change in the scattering function can be observed as a parameter of the system is varied through the critical value at which  $D_c=2N-3$ . Eq. (1) is derived below.

We consider  $N$ -degree-of-freedom, time-independent, open Hamiltonian systems, such that the dy-

<sup>1</sup> Also at Department of Physics.

<sup>2</sup> Also at Department of Electrical Engineering and Physics.

namics is time reversible. That is, if  $x=X(t)$ ,  $p=P(t)$  are solutions of Hamilton's equations (where  $x$  and  $p$  are the  $N$ -dimensional configuration and momentum vectors), then  $x=X(-t)$ ,  $p=-P(-t)$  are also solutions. The dynamics will be reversible if the Hamiltonian is an even function of  $p$ . For example,

$$H = \frac{1}{2}p^2 + V(x). \quad (2)$$

Let  $D_s$  and  $D_u$  denote the dimensions of the stable and unstable manifolds of the chaotic invariant set. Due to the assumed time reversibility of the dynamics, these dimensions must be equal.

$$D_s = D_u. \quad (3)$$

(Non-time-reversible dynamics occurs, for example, when magnetic fields are present and leads to Hamiltonians which are not even functions of  $p$ . In these cases, (3) need not hold.)

We shall be interested in the dimension of intersections of sets lying in the energy surface. As background, we note the following. Let  $S_1$  and  $S_2$  denote two subsets of a  $D_c$ -dimensional manifold, and let their dimensions be denoted  $D(S_1)$  and  $D(S_2)$ . If  $S_1$  and  $S_2$  are smooth surfaces, then generically

$$D(S_1 \cap S_2) = D(S_1) + D(S_2) - D_c. \quad (4)$$

if the right hand side is nonnegative and  $S_1 \cap S_2$  is not empty. If it is negative, then  $S_1$  and  $S_2$  do not have a generic intersection. For example, two one-dimensional lines in a three-dimensional space may intersect at a point, but a slight perturbation of the position of the lines typically removes the intersection. Thus the original intersection is not "generic". We wish to apply (4) also to the case where  $S_1$  is fractal and  $D(S_1)$  is its Hausdorff dimension with noninteger value<sup>\*1</sup>. For this purpose, we refer to the theorems in ref. [4]. As an example of these results, consider the case of a fractal set  $S_1$  lying in a rectangular region of a plane ( $D_c=2$ ). Now randomly choose a straight line  $S_2$  in the plane by first choosing a point with uniform probability distribution in the rectangle and then placing the line through this point at an angle chosen randomly with uniform probability in  $[0, 2\pi]$ . If the left hand side of (4) is neg-

ative (i.e.,  $D(S_1) < 1$  since  $D_c=2$  and  $D(S_2)=1$ ), then the probability that the randomly chosen line intersects the fractal set  $S_1$  is zero. If the right hand side of (4) is positive (i.e.,  $D(S_1) > 1$ ), then there is a positive probability that  $S_1 \cap S_2$  is not empty; and, furthermore, if  $S_1 \cap S_2$  is not empty, then  $D(S_1 \cap S_2)$  is given by (4) with probability one.

We now apply (4) to the chaotic scattering situation. Since the intersection of the stable and unstable manifolds is the chaotic set, we see that (3) and (4) with  $D_s = D_u = 2N-1$  yield

$$D_c = N + d_1, \quad (5)$$

with

$$d_1 = \frac{1}{2}(D_c - 1). \quad (6)$$

We now observe that the fractal set of singular values for the scattering function corresponds to points on the stable manifold of the chaotic set. The orbits originating from such points asymptote to the chaotic set. Orbits originating near these points will spend a long time "bouncing around" in the scatterer before leaving the scattering region; that is, they stay close to the chaotic set for a long time and hence are sensitive to small perturbations of their initial conditions. Let  $d_1$  denote the fractal dimension of the set of singular values of the variable in the scattering function which characterizes the orbit before scattering. Sweeping this single, before-scattering variable corresponds to moving along a curve in the  $D_E$ -dimensional energy surface. Thus  $d_1$  is the dimension of the intersection of the stable manifold of the chaotic set with a one-dimensional set, and (4) yields,  $d_1 = D_s + 1 - D_E$ , or

$$d_1 = d_1 + 2 - N, \quad (7a)$$

$$d_1 = \frac{1}{2}D_c + \frac{3}{2} - N, \quad (7b)$$

where in (7b) we have used (6). (Note that (7a) applies whether or not the Hamiltonian is time reversible, while (6) and hence (7b) require time reversible dynamics.) If the right hand side of (7) is negative, then there is zero "probability" of intersection, and we will typically never observe fractal properties of the scattering function. Requiring  $d_1 > 0$  in (7b) yields the previously stated condition for fractal behavior in the scattering function, eq. (1).

We emphasize that the critical value,  $D_c = 2N-3$ ,

<sup>\*1</sup> Formula (4) applies if  $S_1$  is a Souslin set and  $S_2$  is a smooth surface. A Souslin set is the union of countable intersection of closed sets. See ref. [4].

for observation of fractal behavior in the scattering function results under the assumption that the scattering function is obtained by varying a single before-scattering variable holding all the others fixed. If instead, we choose to consider scattering functions which depend on  $n$  independent before-scattering variables with the others held fixed, then similar considerations can be applied. In this case, fractal behavior in the  $n$ -independent-variable scattering function is typically observable if  $D_c > 2(N-n) - 1$  (for time reversible systems); and the fractal dimension of the set on which the scattering function is singular is  $d_n = d_c + n + 1 - N$ . In such cases we say that the chaotic scattering is an " $n$ -dimensional observable". Since, as a practical matter, it is much easier to examine a function of a single independent variable, we expect the one-dimensional observable case to be of most interest.

We check the above qualitative features in a simple system exhibiting chaotic scattering. It consists of a point particle of unit speed bouncing between four identical hard spheres. The centers of the spheres are located at the vertices of a regular tetrahedron (fig. 1) of unit edge length. The spheres are labeled by  $\{0, 1, 2, 3\}$ . The coordinates of their centers  $\{(x_i, y_i, z_i), i=0, 1, 2, 3\}$  are:

$$(x_0, y_0, z_0) = (0, 0, \sqrt{\frac{3}{4}}),$$

$$(x_1, y_1, z_1) = (\frac{1}{2}, -1/2\sqrt{3}, 0),$$

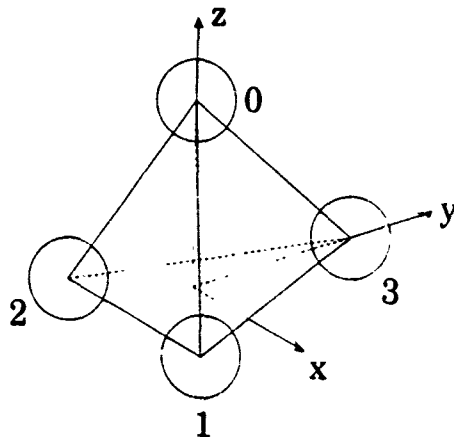


Fig. 1. The geometry of the scatterer, four reflecting hard spheres sitting at the vertices of a regular tetrahedron.

$$(x_2, y_2, z_2) = (-\frac{1}{2}, -1/2\sqrt{3}, 0),$$

$$(x_3, y_3, z_3) = (0, 1/\sqrt{3}, 0).$$

Thus the bottom of the tetrahedron sits on the plane  $z=0$ . The radius of the spheres  $R$  is the only adjustable parameter in the system, and the spheres do not intersect as long as  $R < \frac{1}{2}$ .

There are an infinite number of trapped orbits, periodic or aperiodic, in our system. These orbits are all unstable since small displacements from a trapped orbit are magnified exponentially by the defocusing effect of the spheres. All trapped orbits can be uniquely coded by a bi-infinite sequence  $\{a_i\}$  of four symbols  $\{0, 1, 2, 3\}$  in the following way. We introduce a discrete time as the time of collision of the particle with one of the four spheres. The symbol  $a_i$  is set to  $k$  if the particle collides with sphere  $k$  at time  $i$ . Obviously, the particle cannot hit the same sphere it collided with at the immediately previous time. Therefore, when  $R$  is small enough, the sole constraints on the symbol sequence of trapped orbits is  $a_i \neq a_{i-1}$ . If the symbol sequence is periodic, the corresponding orbit is also periodic. For instance, the orbit bouncing between sphere one and sphere two is of period two, and its symbol sequence is  $\{\dots, 1, 2, 1, 2, \dots\} = [1, 2]$ , where the square bracket denotes the periodicity. There are a total of six period-two orbits:  $[0, 1]$ ,  $[0, 2]$ ,  $[0, 3]$ ,  $[1, 2]$ ,  $[1, 3]$ ,  $[2, 3]$ . There is no period-one orbit due to the constraint  $a_i \neq a_{i-1}$ . The number of trapped periodic orbits grows exponentially with the period. The exponent is the topological entropy of the set of trapped orbits. For our system, when  $R$  is small enough, the topological entropy is  $\log(3)$ .

To proceed, imagine the following situation. We choose a plane below the scattering tetrahedron of spheres,  $z = -K$ ,  $K > R$ . We then consider trajectories originating from initial conditions  $(x_0, y_0)$  on this plane and with initial velocity straight upward (i.e., parallel to the  $z$ -axis). We refer to  $(x_0, y_0)$  as the impact parameters. For all initial conditions  $(x_0, y_0)$ , we define a nonnegative integer valued function  $T(x_0, y_0)$  which we call the time delay function. Its value is given by the total number of collisions with the hard spheres experienced by the particle with impact parameters  $(x_0, y_0)$ . For almost all impact parameters, this function is finite, corresponding to a finite trapping time of the particle in the system.

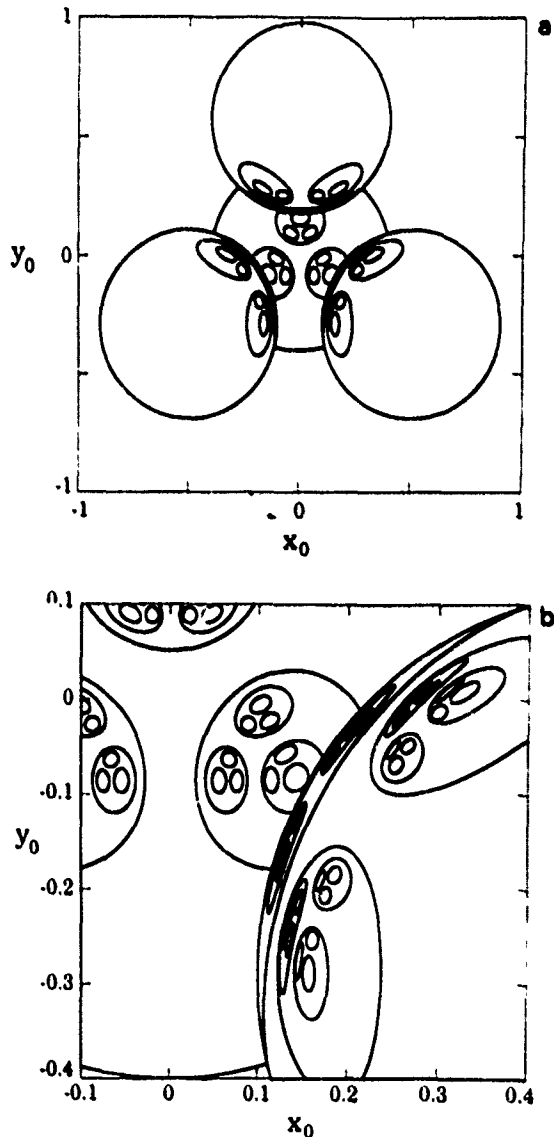


Fig. 2. (a) Hierarchical construction of the Cantor set structure of the stable manifold,  $R=0.4$ ; (b) blowup of (a).

However, there are certain trajectories which remain in the system for an arbitrarily long time. Initial conditions  $(x_0, y_0)$  for these trajectories are distributed on a Cantor set. This Cantor set is the intersection in the five-dimensional energy surface of the stable manifold of the trapped unstable set with the two-dimensional plane  $z = -K$ ,  $p_x = p_y = 0$ . The time delay function is singular on this Cantor set.

To see the Cantor set structure of the stable manifold, we consider the particle trajectories in more

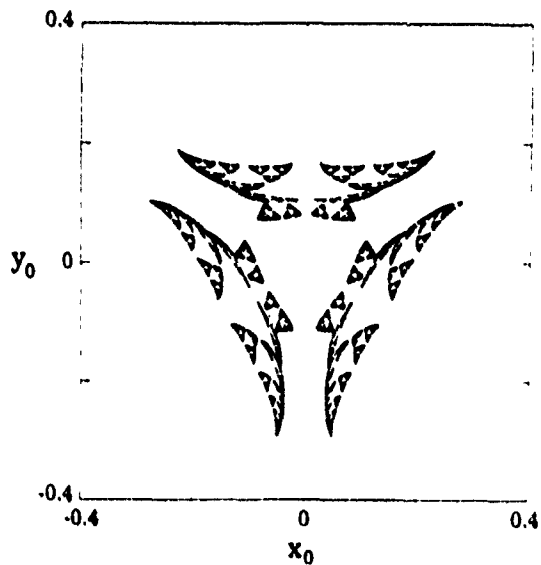


Fig. 3. The intersection of the stable manifold with the hyper-plane plane  $z = -K$ ,  $p_x = p_y = 0$ ,  $R = 0.48$ .

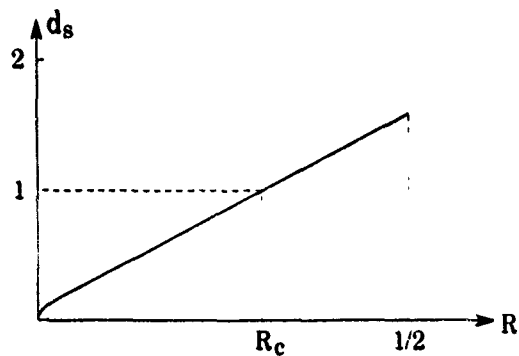


Fig. 4. Schematic illustration of the dimension  $d_s$  as a function of  $R$ .

detail. For some impact parameters, the particle will not hit any of the four hard spheres and will go straight off to infinity. Those initial conditions from which the particle hits one of the four spheres at least once are the vertical projection of the four spheres onto the plane of initial conditions. They are the four big circular disks in fig. 2. We denote this set from which orbits experience at least one bounce by  $C_1$ . Inside each big disk, there are three small deformed disks, from which the particle hits the four spheres at least twice. These are images of the other three

spheres in the mirror of the first sphere. Thus we have a set  $C_2$  of nine small disks from which orbits bounce at least twice. Within each small disk, there are three smaller disks  $C_3$ , from which the particle hits the hard spheres three or more times. The resulting set of this hierarchical disk organization, given by  $\bigcap_{i=1}^{\infty} C_i$ , is the Cantor set illustrated in fig. 3. Starting from any point in this set, the particle bounces between the four hard spheres forever, never escaping to infinity.

The fractal dimension of this Cantor set is  $d_1$ , and

is related to the dimension of the stable manifold  $D_s$  by  $D_s = 3 + d_1$ . It is reasonable to presume that  $d_1$  is a monotonically increasing function of the radius  $R$ . When  $R$  is zero, there is no strange set on the plane of initial conditions, and hence  $d_1$  is zero. For small  $R$ , the dimension  $d_1$  increases sharply with  $R$ ,

$$d_1 \sim 1/\ln(R^{-1})$$

as can be shown by an argument similar to one given in ref. [3]. On the other hand, if  $R \geq 1/\sqrt{3}$ , the region

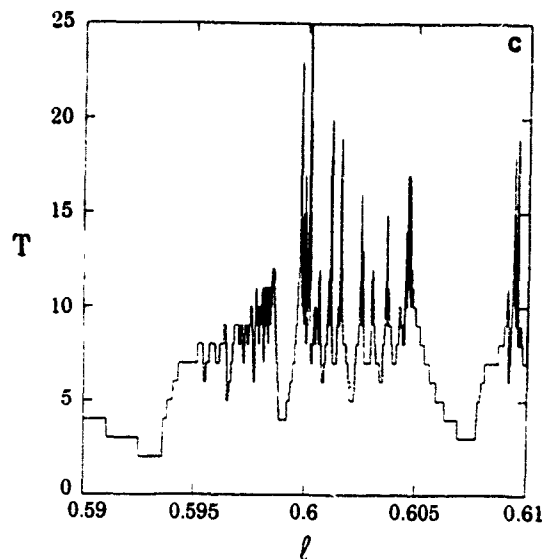
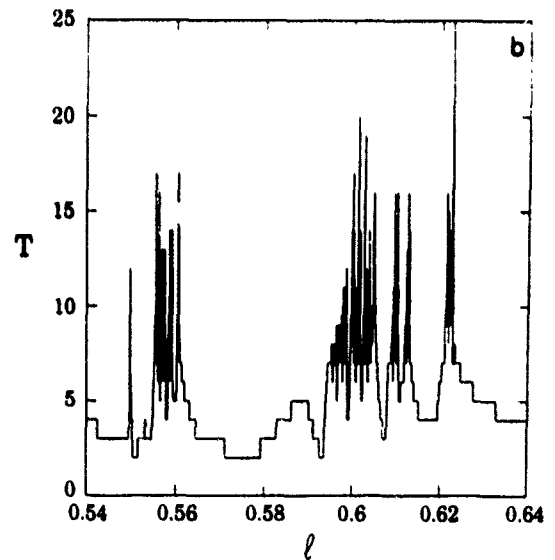
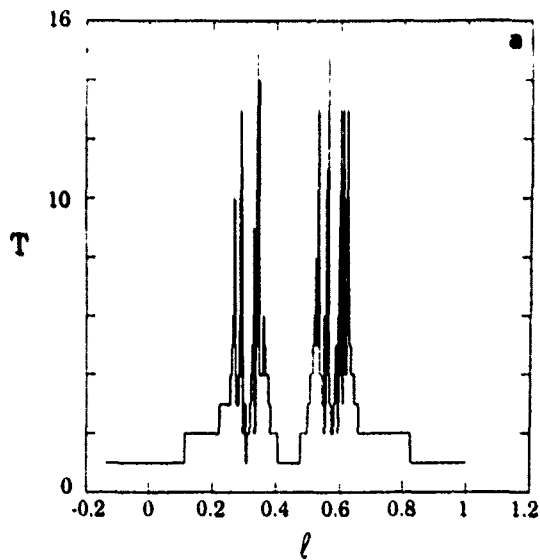


Fig. 5. (a) The time delay as a function of the distance  $l$  along the one-dimensional line cut in a case exhibiting chaotic scattering,  $R=0.48$ ; (b) blowup of (a); (c) blowup of (b).

between the four spheres is closed to the outside (in this case, the spheres intersect since  $R > \frac{1}{2}$ ), and all the points in this closed region are trapped. Hence, all the points in the closed region are on the stable manifold (i.e., the chaotic set and its stable manifold are the same set). The dimension of the stable manifold in this case is equal to the dimension of the energy surface,  $D_s = 5$ , and thus  $d_s = 2$ . Therefore, if we vary  $R$  between 0 and  $1/\sqrt{3}$ , the dimension  $d_s$  increases from 0 to 2. Thus there will be a value  $R = R_c$

at which  $d_s = 1$ , and the scattering will change qualitatively as  $R$  increases through  $R_c$ . Below  $R_c$ , we will not see chaotic scattering from a one-dimensional cut in the plane of initial conditions. A question of prime interest in this context is whether  $R_c < \frac{1}{2}$ . If it is, then we will be able to see chaotic scattering for typical one-dimensional cuts for  $R$  in a range of values ( $R_c < R < \frac{1}{2}$ ) such that the spheres do not intersect.

We used a box counting algorithm to determine the fractal dimension  $d_s$ . We cover the Cantor set

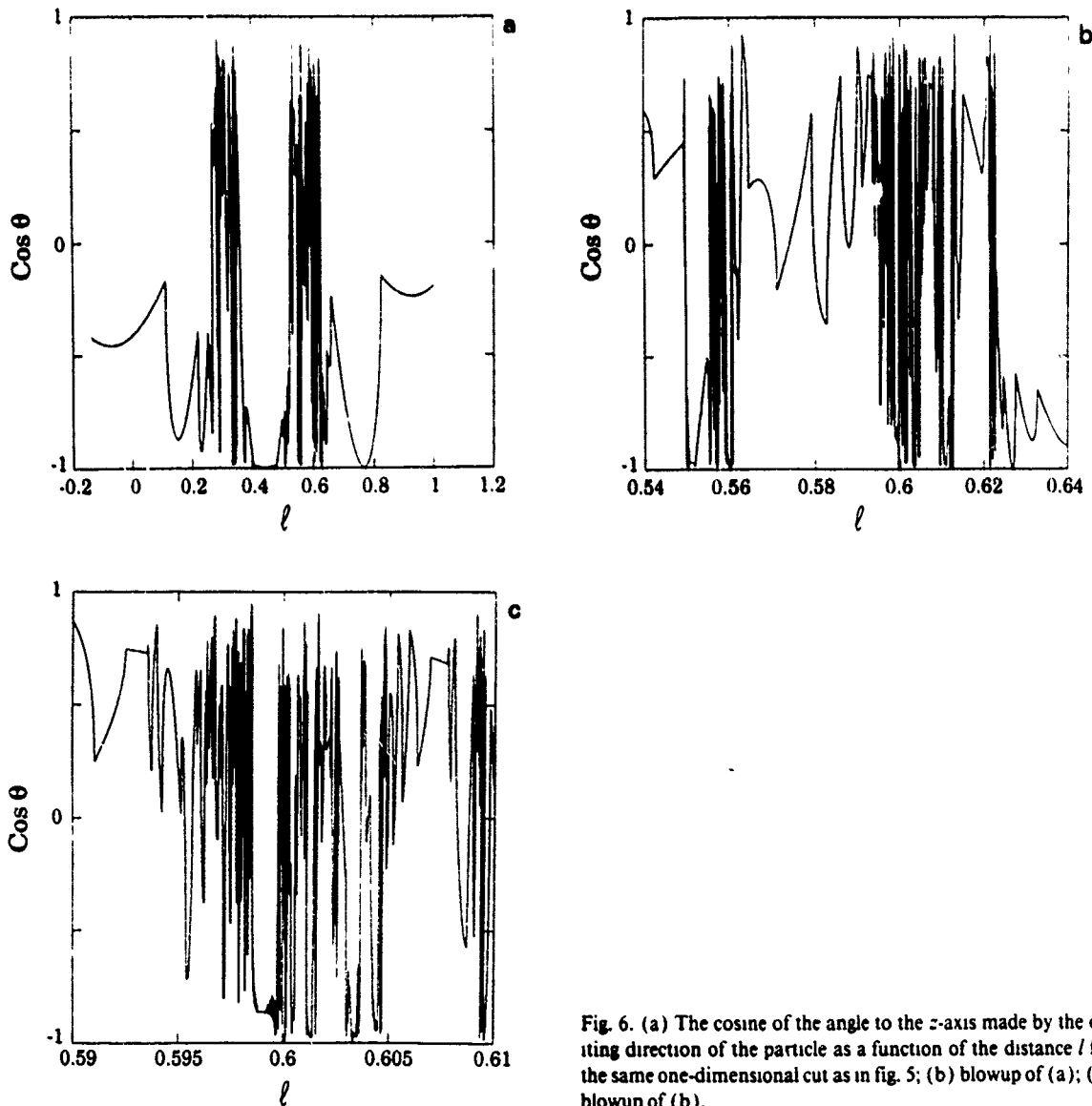


Fig. 6. (a) The cosine of the angle to the  $z$ -axis made by the exiting direction of the particle as a function of the distance  $l$  for the same one-dimensional cut as in fig. 5; (b) blowup of (a); (c) blowup of (b).

generated above by squares of edge length  $\epsilon$ , then in the limit  $\epsilon \rightarrow 0$ , the number of squares  $N(\epsilon)$  needed for the covering scales as

$$N(\epsilon) \approx \epsilon^{-d_1}.$$

The exponent  $d_1$  can be determined by a least-squares fit of  $N(\epsilon)$ . When  $R=0.48$ , we found  $d_1$  is approximately 1.4. Thus we verify the important result that  $R_c < \frac{1}{2}$ . See the schematic illustration in fig. 4. We also computed  $d_1$  at a smaller  $R$  value,  $R=0.4$ , at which we obtain  $d_1 \approx 1.07$ . Using a linear extrapolation from these two computed values of  $d_1$ , we estimate  $R_c \approx 0.38$ .

We now describe some of our numerical results at  $R=0.48$ . Since  $d_1 \approx 1.4 > 1$ , we expect, with positive probability, to see chaotic scattering from a randomly chosen one-dimensional cut in the plane of initial conditions. The fractal dimension of the non-escaping set on this one-dimensional line should be equal to  $d_1 = d_1 - 1 = 0.4$ . We check this by generating one-dimensional random cuts in the plane. We pick a random point in the square centered at the point  $x=y=0$ , of edge length  $2R$ . Then we draw a line at a random angle through this point. Restricting initial conditions to this line, we then plotted the "time delay" (i.e., the total number of bounces from spheres experienced by a particle) as a function of distance  $l$  along this line. Out of thirty such lines, we found nineteen cases exhibiting a fractal set of singularities of the time delay function. A typical form of the time delay function restricted to the one-dimensional line in cases where we observe chaotic scattering is shown in fig. 5a. From the blowups plotted in figs. 5b and 5c, we conclude that the singularities in the time delay function are apparently distributed in a fractal set. Another way to confirm this is to examine the dependence of the scattering function giving the exiting particle direction. Fig. 6a shows plots of the cosine of the angle  $\theta$  to the  $z$ -axis made by the velocity of an exiting particle as a function of distance  $l$  along the same randomly chosen line as was used for fig. 5. In regions near singularities, this function oscillates wildly. Successive blowups of this function (figs. 6b and 6c) show qualitative similarity, again indicating fractal singularities.

To determine the fractal dimension of the set of singularities on a one-dimensional line, we use the following algorithm (the usual box counting method

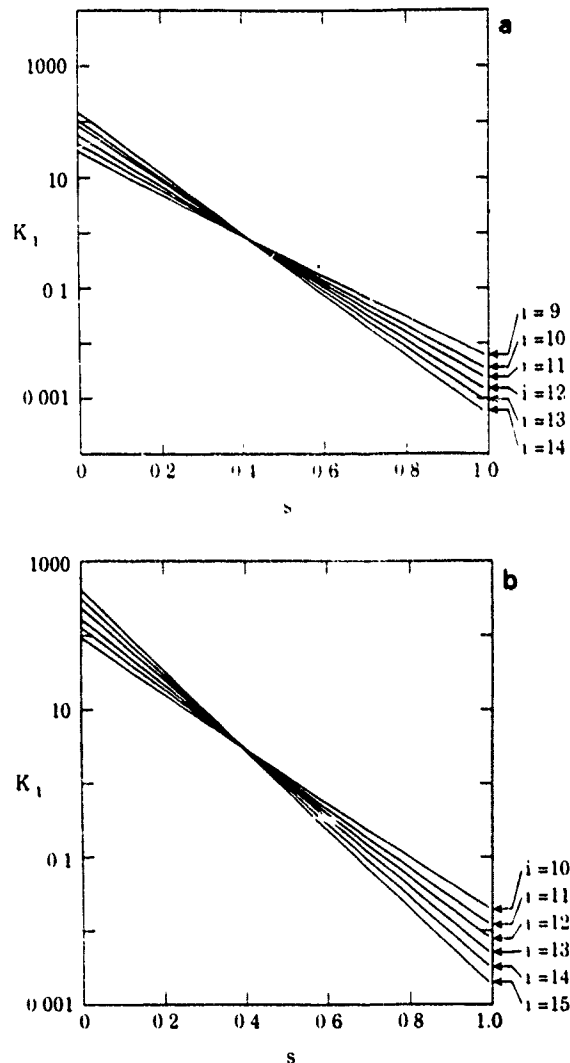


Fig. 7 (a) The Hausdorff sum  $K'(s)$  as a function of  $s$  for different level  $l$ ; (b) the same plot for a different one-dimensional cut

yields an error comparable to the fractal dimension). The time delay function assigns naturally a level structure to the one-dimensional line. At level  $l$ , we measure the length of all the intervals where the time delay function is greater than or equal to  $l$  and denote them by  $I_j^l$ . Then we form the Hausdorff sum

$$K'(s) = \sum_j (I_j^l)^s, \quad (8)$$

where the sum is taken over all intervals at level  $l$ . When  $l$  tends to infinity, this sum should give the Hausdorff  $s$ -dimensional measure [5]. Therefore, it

is infinite when  $s$  is less than the Hausdorff dimension  $d_1$  of the fractal set, and is zero when  $s$  is greater than  $d_1$ . Hence, we expect that for sufficiently large level  $i$ , the sums  $K'(s)$  for different levels will all intersect with each other at approximately the same point  $s=d_1$ , given by the Hausdorff dimension of the one-dimensional fractal set.

For  $R=0.48$ , numerical calculation indeed shows that the sums  $K'(s)$  for large levels all intersect at approximately the same value, thus yielding an approximation to  $d_1$ . Figs. 7a and 7b plot  $K'(s)$  as a function of  $s$  for different  $i$  for two one-dimensional line cuts of the plane of initial conditions. (Small  $i$  data are not shown here, since they do not reflect the fractal property of the singular set.) Within numerical errors, the intersection points are all centered at  $d_1=0.4\pm 0.05$ . This value is also consistent with results obtained for other cuts exhibiting chaotic scattering and is also consistent with our box counting result  $d_1\approx 1.4$ .

When  $R=0.25$ , the fractal dimension  $d_1$  is less than one. Consistent with this, from 100 random line cuts of the plane of initial conditions, we did not see any

fractal behavior in the scattering function.

We thank Ittai Kan for discussion. This work was supported by the Office of Naval Research (Physics), by the Department of Energy (Basic Energy Sciences) and by the Advanced Research Projects Agency.

## References

- [1] B. Eckhardt, *Physica D* 33 (1988) 89;  
U. Smilansky, The classical and quantum theory of chaotic scattering, Lectures at Les Houches, Session LII, Chaos and quantum physics, eds. M.-J. Giannoni, A. Voros and J. Zinn-Justin (Elsevier, Amsterdam, 1990), to be published.
- [2] R. Blümel and U. Smilansky, *Phys. Rev. Lett.* 60 (1988) 477;  
P. Gaspard and S. Rice, *J. Chem. Phys.* 90 (1989) 2242, 2255;  
P. Cvitanović and B. Eckhardt, *Phys. Rev. Lett.* 63 (1989) 823.
- [3] S. Bleher, E. Ott and C. Greb, *Phys. Rev. Lett.* 63 (1989) 919.
- [4] P. Mattila, *Acta Math.* 152 (1984) 77; *Ann. Acad. Sci. Fennicae A 1* (1975) 227.
- [5] K.J. Falconer, *The geometry of fractal sets* (Cambridge Univ. Press, Cambridge, 1985).

## Cross-sections of chaotic attractors

Qi Chen and Edward Ott<sup>1</sup>

*Laboratory for Plasma Research, University of Maryland, College Park, MD 20742, USA*

Received 26 March 1990; accepted for publication 18 May 1990

Communicated by A.P. Fordy

We present an efficient algorithm for constructing cross-sections of chaotic attractors. The technique is particularly useful for studying the structure and fractal dimension of higher dimensional attractors.

One of the central topics in nonlinear dynamical systems theory is the study of the structure and organization of invariant sets under the dynamics. In particular, the geometry of strange attractors [1] is of particular interest. For such studies, the visualization of the strange attractor is important for revealing structure as well as characterizing the attractor. This presents problems when higher dimensional attractors are encountered. For example, the projection of an attractor whose fractal dimension is greater than two to a plane yields a fuzzy blob. Questions such as whether the local structure of a typical higher dimensional strange attractor is the product of a continuum with a Cantor set [2] or is more complex than this cannot be answered by simply taking a projection of the attractor. In addition, numerical determination of the dimension of higher dimensional fractal sets by box-counting algorithms can require enormous memory storage and CPU time. If feasible, taking *cross-sections* of the attractor (i.e., intersections of the attractor with a surface) might offer a way of both elucidating the geometry of the attractor and of estimating its dimension.

In this regard, two procedures for taking a cross-section of a chaotic attractor were proposed by Lorenz [2], and the first of them was extended and further developed by Kostelich and Yorke [3]. This latter procedure is basically as follows. An orbit on the

chaotic attractor is followed until it comes near the desired cross-section plane. Through a subsidiary calculation, a local approximation to the unstable manifold through that point is found. Then the intersection of the approximate unstable manifold and the desired cross-section plane is determined, thus projecting the orbit point onto the cross-section plane. Assuming the attractor is smooth in the unstable direction (or directions), this intersection approximates to a point in the cross-section of the attractor. Repeating this procedure many times as an orbit is followed, a cross-section picture of the attractor is built up.

In this note, we consider Lorenz's second procedure for taking numerical cross-section. Compared to the first procedure, this procedure can be easier to implement and yield faster computer computation. On the other hand, the method has certain limitations which will be discussed. Consider an  $N$ -dimensional invertible map,  $x_{n+1} = F(x_n)$ . Choose a compact volume  $V$  which contains the chaotic attractor. We shall find the cross-section of an  $m$ -dimensional hyperplane with the unstable manifolds of the invariant sets contained in  $V$ . This will typically include the attractor. By inverting the map, the attractor becomes a repeller. Consider a point  $x$  in  $V$  and examine its preimages  $F^{-1}(x)$ ,  $F^{-2}(x)$ , ...,  $F^{-n}(x)$ . Let  $T(x)$  denote the smallest value of  $n$  such that  $F^{-n}(x)$  is not in  $V$ . We call  $T(x)$  the inverse escape time from  $V$ . Under the inverse map, all points in the region  $V$  will finally escape except for those on the unstable manifolds of the invariant sets con-

<sup>1</sup> Also at Department of Electrical Engineering and Department of Physics

tained in  $V$ . This set, of course, includes the repeller of the inverse map originating from the chaotic attractor of the forward map. Points on the unstable manifolds of the invariant sets in  $V$  correspond to singular points of the inverse escape time function ( $T(x)=\infty$ ). (We assume that the inverse map has no attractors in  $V$ . For example, the inverse of a strictly contractive map (e.g., the Hénon map) can have no attractors.) Thus, if we start initial conditions from a hyperplane and collect all the singular points of the inverse escape time function of the map, we find the intersection of the hyperplane with the unstable manifolds of the invariant sets in  $V$ , and this typically includes the cross-section of the attractor with the hyperplane. In practice, we do not determine the singular points but rather we determine a succession of nested sets containing the singular points. We do this by computing  $x$  values for which  $T(x) \geq N$  for successively larger values of  $N$ . To obtain the intersection with the attractor, one should reject points which satisfy  $T(x) \geq N$  but do not lie approximately on the attractor. In principle, this can be done by calculating the Lyapunov exponents (or other ergodic quantities) of  $F^{-1}$  for each  $x$  satisfying  $T(x) \geq N$  along the orbit  $x, F^{-1}(x), \dots, F^{-(N-1)}(x)$ . For large  $N$ , these exponents will approximate the negatives of the Lyapunov exponents of the forward map on the attractor, provided that  $x$  lies approximately on the attractor. If  $x$  does not lie approximately on the attractor, then the Lyapunov exponents for the inverse map starting from  $x$  will approximate those for another invariant set in  $V$  and will differ substantially from the exponents of the attractor. In this case the point  $x$  is rejected. It will not always be possible to apply this Lyapunov exponent test, because  $N$  must be sufficiently large to obtain reliable estimates of the Lyapunov exponents of the inverse map. Alternatively, one can omit the Lyapunov exponent test altogether. In this case, the set obtained may be larger than that for the attractor. Thus a calculation of the fractal dimension of this set yields an upper bound for the fractal dimension of the attractor. In our numerical examples, we have not applied the Lyapunov exponent test. Nonetheless, as shown below, for these examples, the method appears to yield very good approximations to the actual attractor, and the calculated dimen-

sions agree with previous calculations which resolved only the attractor set.

The dimension of the intersection set in the cross-section plane is related to the dimension of the unstable manifold set by a result of Mattila [6]. If the Hausdorff dimension  $D$  of a bounded fractal set lying in an  $N$ -dimensional space is greater than  $N-m$ , then a random cut by an  $m$ -dimensional hyperplane intersects the set with positive probability; if it does intersect the fractal set, the dimension  $d$  of the intersection set is related to  $D$  by

$$D = d + (N - m) \quad (1)$$

with probability one. Hence, by generating the cross-section of the attractor and measuring the dimension of the cross-section set, we determine the dimension of the strange attractor.

To illustrate our algorithm, we first calculate one-dimensional cross-sections of the Hénon attractor. The Hénon attractor is generated by the following map,

$$x_{n+1} = a - x_n^2 + by_n, \quad y_{n+1} = x_n. \quad (2)$$

At parameter values  $a=1.4, b=0.3$ , Hénon observed that there exists a chaotic attractor. Numerical box counting techniques for the calculation of the dimension of a strange attractor were first applied by Russell et al. [4], who obtained a result for the dimension of the Hénon attractor. A more accurate result was obtained by Grassberger who found that the capacity dimension is approximately  $1.28 \pm 0.01$  [5]. However, from different least squares fits of the slope, the dimension takes values between 1.22 and 1.30.

Fig. 1 shows the Hénon attractor. It can be shown that the attractor is included in the square  $[-2.0, 2.0] \times [-2.0, 2.0]$ . This is the region  $V$  which we use for calculating the inverse escape time function. We take a horizontal one-dimensional cross-section through the point  $x=0, y=0$  and calculate  $T(x)$  at regularly spaced intervals along this line. This is shown in fig. 2a. We see there is a natural Cantor set level structure in the inverse escape time function. At level 0, there is one interval from which it requires at least one backward iterate to escape the square; at level 1, there are two intervals from which it requires at least two backward iterates to escape the square; etc. The intersection of all these intervals is the cross-section of the Hénon attractor. Fig. 2b

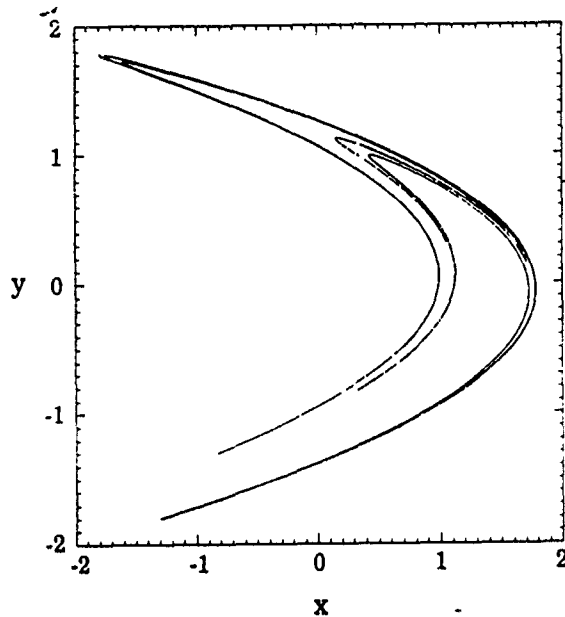


Fig. 1. The Hénon attractor.

shows the same function for the vertical cross-section through the same point  $x=0$ ,  $y=0$ .

To get the fractal dimension of these cross-section sets, we use the following procedure. We denote the lengths of the intervals at level  $i$  by  $l'_i$ . Then we form the Hausdorff sum

$$K'(s) = \sum_i (l'_i)^s, \quad (3)$$

where the sum is taken over all intervals at level  $i$ . When  $i$  tends to infinity, this sum is the Hausdorff  $s$ -dimensional measure [7]. Therefore, it is infinite when  $s$  is less than the Hausdorff dimension  $d$  of the fractal set, and is zero when  $s$  is greater than  $d$ . Hence, we expect that for large  $i$ , the sums  $K'(s)$  versus  $s$  for different levels will intersect with each other at approximately the same point  $s=d$  given by the Hausdorff dimension of the one-dimensional fractal set<sup>†1</sup>. In fig. 3, we show results for the Hausdorff sums for different levels for a typical one-dimensional cut. The lines for this case have intersections in the range

<sup>†1</sup> The numerical application of the Hausdorff sum (3) to find the fractal dimension has been previously used to study chaotic scattering [8]. Results of Nusse and Yorke [9] guarantee that for hyperbolic horseshoes, an interval with successive nested increasing  $T(x)$  contains a point where  $T(x) = \infty$ .

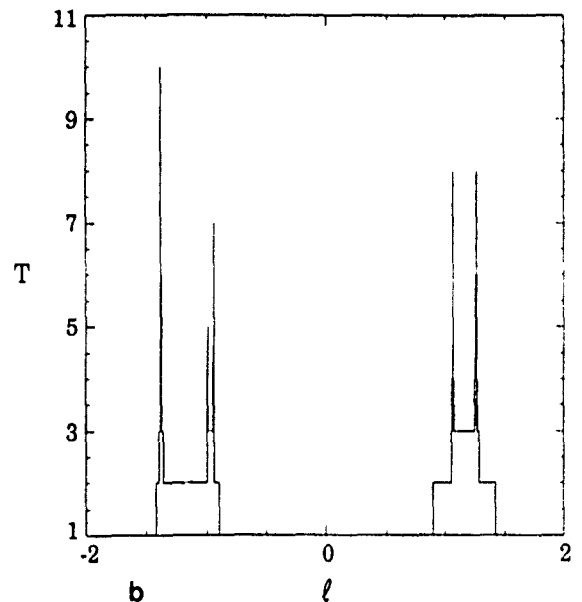
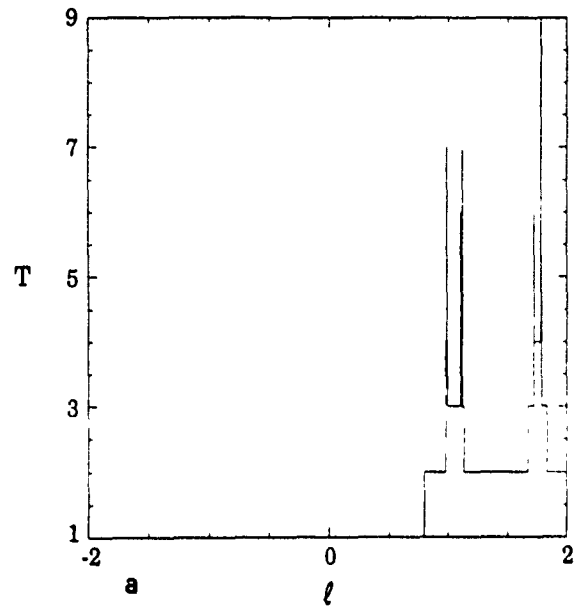


Fig. 2. Inverse escape time function for the Hénon map. (a) Horizontal cut through  $x=0$ ,  $y=0$ . (b) Vertical cut through  $x=0$ ,  $y=0$ .

$d \approx 0.24$  to  $0.30$ . Examining many different one-dimensional horizontal and vertical cuts, we estimate  $d$  to lie in the range  $0.20$  to  $0.34$ . From formula (1), the dimension of the Hénon attractor is approximately  $D \approx 1.20$ – $1.34$ . The whole calculation for a

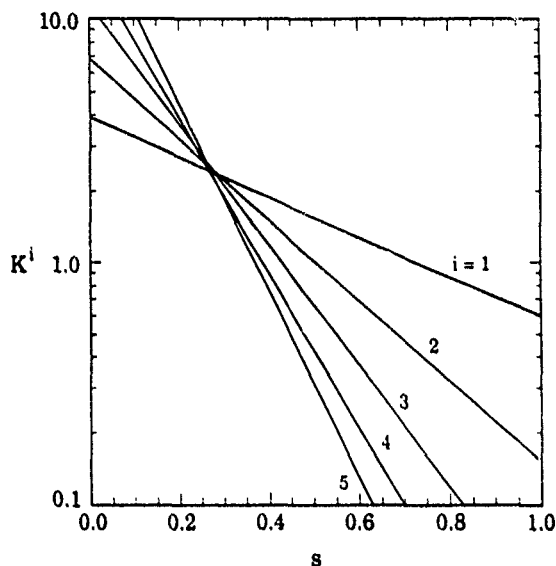


Fig. 3. The Hausdorff sum  $K^i(s)$  as a function of  $s$  for different levels  $i$  for the one-dimensional vertical cut through  $x=0.8$ ,  $y=0.0$ .

cut involved very little computer memory and took less than 5 seconds on the Cray XMP computer.

Our second example is the double rotor attractor generated by the following four-dimensional, volume-contracting map [10],

$$\begin{aligned} \begin{pmatrix} x_1^{n+1} \\ x_2^{n+1} \end{pmatrix} &= M_1 \begin{pmatrix} y_1^n \\ y_2^n \end{pmatrix} + \begin{pmatrix} x_1^n \\ x_2^n \end{pmatrix} \bmod 1, \\ \begin{pmatrix} y_1^{n+1} \\ y_2^{n+1} \end{pmatrix} &= M_2 \begin{pmatrix} y_1^n \\ y_2^n \end{pmatrix} + \begin{pmatrix} (c_1/2\pi) \sin(2\pi x_1^{n+1}) \\ (c_2/2\pi) \sin(2\pi x_2^{n+1}) \end{pmatrix}. \end{aligned} \quad (4)$$

Here  $x_1, x_2$  take values from the unit interval  $[0, 1]$ , and  $y_1$  and  $y_2$  take values from the real line. At parameter values given by

$$M_1 = \begin{pmatrix} -5.8 & -6.602 \\ -6.602 & -12.40 \end{pmatrix},$$

$$M_2 = \begin{pmatrix} 0.7496 & 0.1203 \\ 0.1203 & 0.8699 \end{pmatrix},$$

$$c_1 = 0.3536, \quad c_2 = 0.5,$$

Kostelich and Yorke [3] find that there is a chaotic attractor. Since the two  $x$ -directions of the double rotor map are compact, we choose for  $V$  the hypercube box given by  $\max(|y_1|, |y_2|) \leq y_{\max}$ . Starting from a uniform distribution of initial points in the

cross-section plane, we collect those points from which after some chosen maximum number of iterates of the inverse map  $n_{\max}$ , the point remains in the hypercube region. Fig. 4 shows two two-dimensional cross-sections of the attractor using our algorithm ( $y_{\max}=0.5$  and  $n_{\max}=15$ ). The pictures in fig. 4 appear to be identical to those in ref. [3].

To find the fractal dimension of the chaotic at-

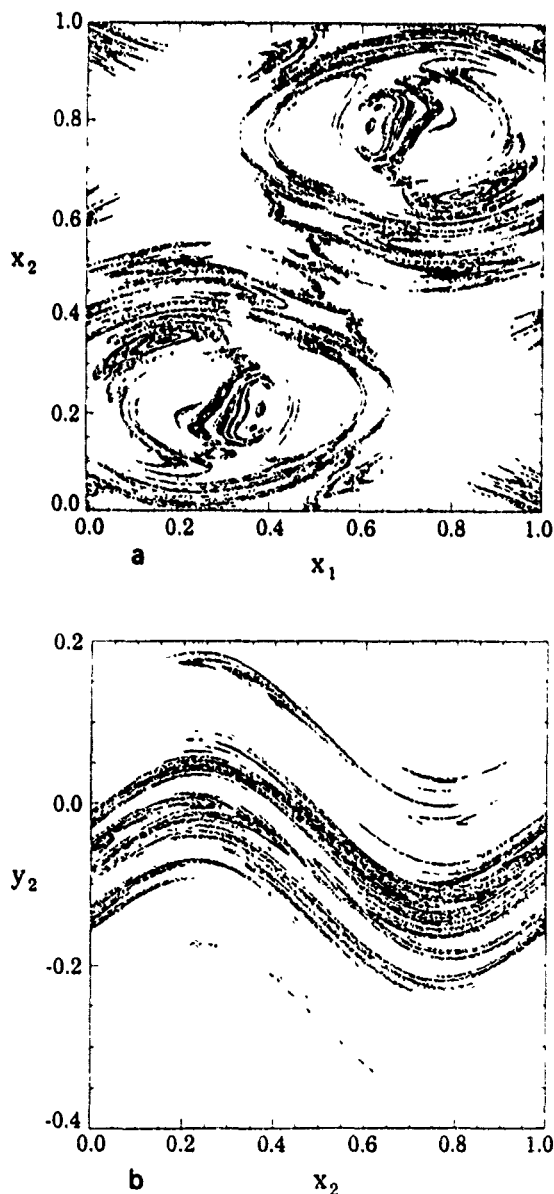


Fig. 4. Cross-sections of the double rotor attractor. (a) Cross-section at  $y_1=0$ ,  $y_2=0$ . (b) Cross-section at  $y_1=0$ ,  $x_1=2/2\pi$ .

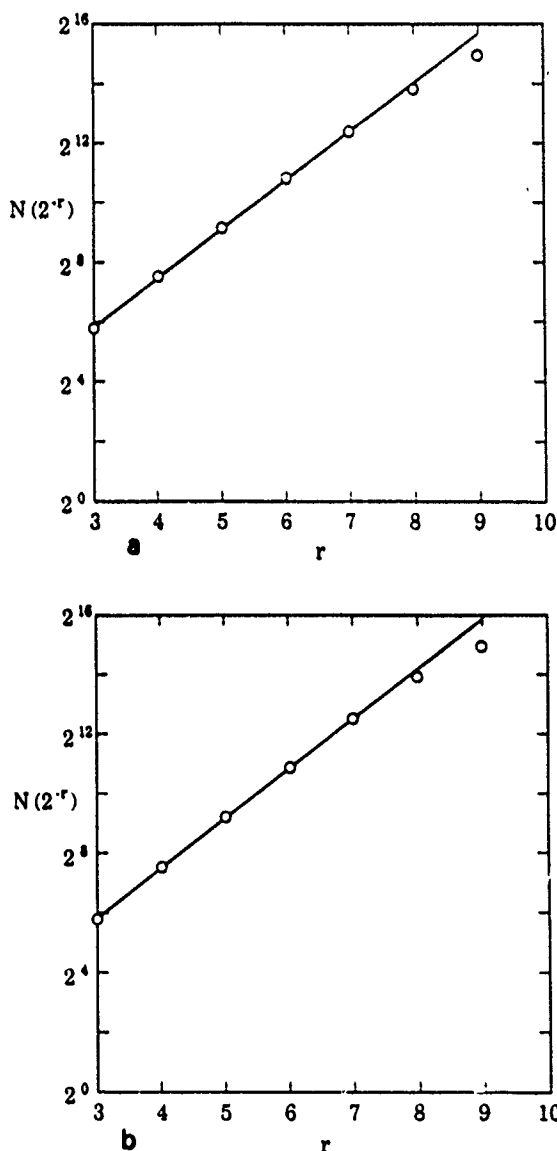


Fig. 5. (a)  $N(\epsilon)$  as a function of  $\epsilon$  for the cross-section set in fig. 4a. The least squares fit gives a capacity dimension  $d = 1.67 \pm 0.05$ . (b) The same plot for fig. 4b. The least squares fit gives  $d = 1.63 \pm 0.05$ .

tractor, we used a box counting algorithm. We cover the resulting cross-section set with squares from a grid of edge length  $\epsilon$ . In the limit  $\epsilon \rightarrow 0$ , the number of

squares  $N(\epsilon)$  needed for the covering scales as

$$N(\epsilon) \sim \epsilon^{-d}. \quad (5)$$

The exponent  $d$  is determined by a least squares fit of a straight line to a log-log plot of  $N(\epsilon)$ . In fig. 5, we calculate the capacity dimension  $d$  for the cross-section sets of figs. 4a and 4b. The two values of  $D = d + 2$  determined from least squares fitting are 3.67 and 3.63. According to the estimates of ref. [3], the information dimension lies in the range 3.61 to 3.68. Thus we find that the values of the capacity and information dimensions (the latter must be smaller) are apparently quite close to each other.

In conclusion, we have presented an efficient algorithm for calculating cross-sections of strange attractors. This method may be useful for the estimation of the fractal dimension of higher dimensional chaotic attractors.

We acknowledge helpful conversations with Mingzhou Ding and James Yorke. This work was supported by the Office of Naval Research (Physics), by the Department of Energy (Basic Energy Sciences) and by the Advanced Research Projects Agency. The computation was done at the National Energy Research Supercomputer Center.

## References

- [1] J.D. Farmer, E. Ott and J.A. Yorke, *Physica D* 7 (1983) 153.
- [2] E.N. Lorenz, *Physica D* 13 (1984) 90; 17 (1985) 279.
- [3] E. Kostelich and J.A. Yorke, *Physica D* 24 (1987) 263.
- [4] D.A. Russell, J.D. Hanson and E. Ott, *Phys. Rev. Lett.* 45 (1980) 1175.
- [5] P. Grassberger, *Phys. Lett. A* 97 (1983) 224.
- [6] P. Mattila, *Acta Math.* 152 (1984) 77, *Ann. Acad. Sci. Fenn. A* 1 (1975) 227.
- [7] K.J. Falconer, *The geometry of fractal sets* (Cambridge Univ. Press, Cambridge, 1985).
- [8] Q. Chen, M. Ding and E. Ott, *Phys. Lett. A* 145 (1990) 154.
- [9] H.E. Nusse and J.A. Yorke, *Physica D* 36 (1989) 137.
- [10] C. Grebogi, E. Kostelich, E. Ott and J.A. Yorke, *Physica D* 25 (1987) 347.

## Rigorous verification of trajectories for the computer simulation of dynamical systems†

Tim Sauer‡ and James A Yorke§

‡ Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030, USA

§ Institute of Physical Science and Technology, University of Maryland, College Park, MD 20742, USA

Received 30 July 1990, in final form 29 January 1991

Accepted by J D Farmer

**Abstract.** We present a new technique for constructing a computer-assisted proof of the reliability of a long computer-generated trajectory of a dynamical system. Auxiliary calculations made along the noise-corrupted computer trajectory determine whether there exists a true trajectory which follows the computed trajectory closely for long times. A major application is to verify trajectories of chaotic differential equations and discrete systems. We apply the main results to computer simulations of the Hénon map and the forced damped pendulum.

AMS classification scheme numbers: 58F13, 58F15, 65G05, 65L70

### 1. Introduction

Are numerical studies of chaotic systems reliable? More specifically, do computer trajectories 'correspond' to actual trajectories of the system under study? The answer is sometimes no. In other words, there is no guarantee that there exists a true trajectory that stays near a given computer-generated numerical trajectory.

The question is especially pivotal for chaotic systems. Chaotic trajectories exhibit sensitive dependence on initial conditions: two trajectories with initial conditions that are extremely close tend to diverge exponentially from one another. At the same time, a great deal of phenomenological research on chaotic systems relies heavily on computer simulation.

Therefore, the use of an ODE solver on a finite-precision computer to approximate a trajectory of a chaotic dynamical system leads to a fundamental paradox. Because of sensitive dependence on initial conditions, a small truncation or rounding error made at any step during the computation will tend to be greatly magnified by future evolution of the system. Under what conditions will the computed trajectory be close to a true trajectory of the model?

Consideration of simple examples of nonlinear maps illustrate that there are critical points of trajectories where round-off error or other noise can introduce new behaviour. We discuss typical examples in section 2. At such 'glitches' the true trajectories all

† Research supported by the Applied and Computational Mathematics Program of DARPA.

diverge from the numerical trajectory. In this case, there will be no true trajectory that stays near the numerical trajectory. In other cases, the numerical trajectory can be *shadowed*: some true trajectory remains close to the numerical trajectory.

In the present work we state a result (theorem 3.3) which says that if certain quantities evaluated at points of the computer-generated trajectory, called a *pseudo-trajectory*, are not too large, then there exists a true trajectory near the computer-generated one. Rigorous upper bounds for these quantities can be generated by the computer as it produces the pseudo-trajectory. If these quantities satisfy the hypotheses of the theorem, which again can be rigorously checked by the computer, the result is a computer-assisted proof of the existence of a true trajectory near the computer-generated pseudo-trajectory. For example, if the one-step errors in the pseudo-trajectory occur in the tenth decimal place, then the true trajectory that results from the theorem differs from the computer-generated trajectory in approximately the fifth decimal place. In particular, the initial point of the true trajectory can differ from the initial condition of the pseudo-trajectory at most in the fifth decimal place.

A typical application of the theorem is to the forced damped pendulum

$$\ddot{y} + a\dot{y} + \sin y = b \cos t.$$

Setting the parameters  $a = 0.2$  and  $b = 2.4$ , we prove the existence of an apparently chaotic trajectory with initial conditions  $y(0) = \dot{y}(0) = 0$  for time  $t$  ranging from  $t = 0$  to  $t = 10^4\pi$ . This trajectory, for all  $0 \leq t \leq 10^4\pi$ , lies within  $10^{-9}$  of an explicit computer-generated (noisy) trajectory produced with a one-step error of  $10^{-18}$ . There are similar results for other initial conditions and other choices of  $a$  and  $b$ .

To describe the theorem, we make a distinction between discrete and continuous models. Computational methods for approximating trajectories of systems of ordinary differential equations work by a series of small, discrete steps. We can therefore consider computer simulation of discrete systems and autonomous differential equations at the same time if we define a dynamical system to be an invertible map  $f$  on  $R^m$ . (We actually define dynamical system a little more generally, as a sequence of maps  $\{f_n\}$  on  $R^m$ , to also cover the non-autonomous differential equations case.) We will try to keep this distinction clear by using the word *trajectory* for continuous systems and *orbit* for discrete systems.

Consider then a  $\delta$ -pseudo-orbit of a discrete system  $f$ , which we can imagine having resulted from applying a one-step quadrature method with truncation error  $\delta$  to a system of differential equations on  $R^m$ ,  $m \geq 2$ . Assume that we have subspaces  $S_n$  and  $U_n$  at each point  $x_n$  of the pseudo-orbit, which are self-consistent with tolerance  $\delta$ . By this we mean that  $S_n$  and  $U_n$  are complementary subspaces of the tangent space  $R^m$  at  $x_n$  (see figure 1), that unit vectors in  $U_n$  are mapped by  $f$  to within  $\delta$  of  $U_{n+1}$ , and similarly for  $S_n$ . Define the positive number  $r_n$  to be an upper bound for the expansion rate of the linearization  $Df$  along  $S_n$ , and  $t_n$  to be an upper bound for the expansion rate of  $Df^{-1}$  along  $U_n$ . See section 3 for precise definitions.

The quantities which need to be measured to assure the existence of a nearby true orbit are most easily expressed as recurrence relations. Set up a recurrence relation  $C_n$  by beginning with  $C_0 = 0$ , and recursively defining  $C_n = \csc \theta_n + r_{n-1}C_{n-1}$ , where  $\theta_n$  is the angle between  $S_n$  and  $U_n$ . Define  $D_n$  similarly:  $D_N = 0$ , where  $N$  is the length of the pseudo-orbit, and  $D_n = \csc \theta_n + t_n D_{n+1}$  for  $n < N$ . Then as long as the quantities  $C_n$  and  $D_n$  are not too large for all  $n$ , there is a true orbit of  $f$  near the pseudo-orbit. More precisely:

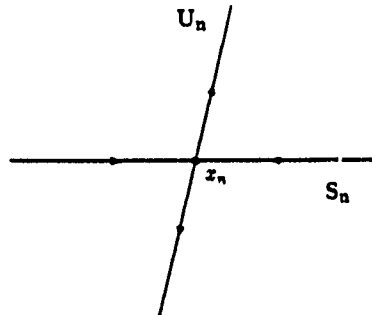


Figure 1. The splitting of the tangent space at the  $n$ th point of the pseudo-orbit.

**Theorem 3.3.** Assume  $\delta < 1/20m^2$  and let  $B$  be a bound on the first and second partial derivatives of  $f$  and  $f^{-1}$ . If

$$\max\{C_n, D_n\} \leq \frac{1}{m^{5/2} B^2 \sqrt{\delta}}$$

for all  $n = 0, \dots, N$ , then there exists an orbit  $\{w_n\}$  of  $f$  such that  $|x_n - w_n| < \sqrt{\delta}$  for  $n = 0, \dots, N$ .

Note that we do not need to assume uniform contraction and expansion along the directions  $S_n$  and  $U_n$ . In other words,  $r_n$  and  $t_n$  do not need to be less than one for all  $n$ .

The proof of the theorem is constructive, in the sense that it uses a procedure for refining noisy orbits originally given in [6]. The essential point of the proof is to show that under the conditions of the theorem, the iterated application of the refinement procedure, beginning with the pseudo-orbit, results in a sequence of refined pseudo-orbits with decreasing noise level, and whose limit is a true orbit. In addition, the true orbit is not too far from the original pseudo-orbit.

The proof can also be considered a justification for using the refinement process computationally on the actual noisy orbit to reduce noise to near machine-precision, but that is a separate issue from the main question we are answering here. This direction is taken up in [7].

A true orbit that stays near the pseudo-orbit is said to *shadow* the pseudo-orbit. Several years ago, Anosov and Bowen proved shadowing results for hyperbolic maps on a differential manifold. The conclusion of Anosov [1] for a hyperbolic map says that, given any prescribed shadowing distance  $\varepsilon$  (between the pseudo-orbit and true orbit) there exists a  $\delta > 0$  so that any  $\delta$ -pseudo-orbit can be  $\varepsilon$ -shadowed by a true orbit. Bowen [2] showed that the same result holds if the map is required only to be hyperbolic on a basic set containing the orbit. Other proofs have been given, and one more is a consequence of the present work.

There are two factors that make the approach of Anosov and Bowen impractical for use in computer experiments. First, the  $\delta$  that is produced can be orders of magnitude smaller than the machine epsilon of existing digital computers. Second, most interesting dynamical systems currently being studied are not hyperbolic.

Theorem 3.3 does not assume that the dynamical system is hyperbolic. Our approach is to prove that as long as the system is sufficiently hyperbolic along the (finite length) numerical trajectory, then that piece of the numerical trajectory can be

shadowed by a true trajectory. On the other hand, when  $f$  is hyperbolic,  $C_n$  and  $D_n$  stay uniformly bounded for all iterates  $n$ , in which case arbitrarily long shadowing trajectories are constructed by the theorem for sufficiently small  $\delta$ . Thus the shadowing theorem of Anosov and Bowen is a consequence of theorem 3.3, as is noted in [9].

In [5, 6] a method is developed which creates computer-assisted proofs of the existence of finite length shadowing orbits on a case-by-case basis. In two dimensions, a small parallelogram is constructed near each point of the numerical orbit in such a way that there is a guarantee of a true orbit whose  $n$ th point lies in the  $n$ th parallelogram. They apply the method to one-dimensional maps and the two-dimensional Hénon and Ikeda maps, none of which are hyperbolic. These papers use auxiliary calculations in 96-bit precision to verify that there are true orbits near the pseudo-orbit, which was produced in 48-bit precision.

The advantage of the present method over [6] is that the auxiliary calculations can now be done in the same precision in which the orbit was calculated. For the maps mentioned above, only 48-bit precision is needed to verify the existence of a pseudo-orbit produced in 48-bit precision.

This fact is especially important when attempting to shadow differential equations. We found that the methods of [6] were not practical, at least for the differential equations we tried. For example, in order to produce long shadowable pseudo-trajectories for the forced damped pendulum, we needed to use a one-step error of no more than  $10^{-18}$ , which already requires 96-bit precision. In this case, there is no extra precision available for the auxiliary calculations of [6].

Thus the new method, superior even for maps, is evidently essential for shadowing differential equations. The improvement is largely gained by sublimating the refinement process, done explicitly in a computer-aided proof in [6], into the proof of theorem 3.3. It is proved here that under the hypotheses of the theorem, the refinement process, when iterated, theoretically converges to a true trajectory.

The main result of this paper was announced in [9], in a slightly less streamlined form. Other work along these lines for the one-dimensional case is reported in [3].

In the next section, it is shown by example that shadowing can fail for some pseudo-trajectories. The details of the main theorem (theorem 3.3) are presented in section 3. Section 4 consists of a number of remarks relevant to the implementation of the computer algorithm based on theorem 3.3. Examples are given in section 5, and section 6 contains the proof of the main theorem.

## 2. Why shadowing works

What makes it possible to find a true orbit near a pseudo-orbit in the presence of sensitive dependence on initial conditions? The short answer is hyperbolicity along the pseudo-orbit. Even for a non-hyperbolic dynamical system, as long as the pseudo-orbit avoids areas of phase space that lack hyperbolicity, it may be possible to find a nearby true orbit. Of course, on typical ergodic chaotic attractors, this avoidance is only done as a matter of degree. Roughly speaking, the pseudo-orbit must stay far away from non-hyperbolic areas compared with the size of the errors being made. Our method essentially relies on measuring how successful the trajectory is in staying hyperbolic.

As a simple example, imagine a map which contracts distances. Assume that the distance between any two points  $x$  and  $y$  is decreased by a factor of  $K$  by the map  $f$ , where  $0 < K < 1$ . Thus  $|f^n(x) - f^n(y)| \leq K^n|x - y|$ . It follows that any pseudo-orbit

can be shadowed by the true orbit beginning at its own initial condition. All distances are contracted, including errors that are made along the pseudo-orbit.

To be more precise, let a  $\delta$ -pseudo-orbit be denoted by  $\{x_0, x_1, \dots, x_N\}$ . Then  $|f(x_0) - x_1| < \delta$ , and further

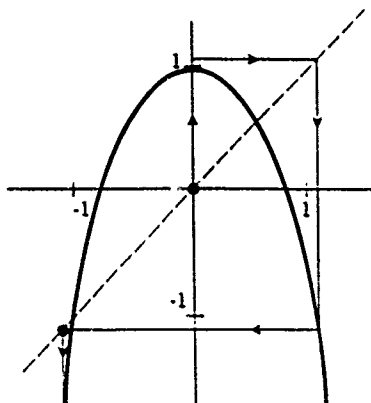
$$\begin{aligned} |f^2(x_0) - x_2| &\leq |f^2(x_0) - f(x_1)| + |f(x_1) - x_2| \\ &\leq K|f(x_0) - x_1| + |f(x_1) - x_2| \\ &\leq (K + 1)\delta. \end{aligned}$$

Continuing in this way,  $|f^n(x_0) - x_n| \leq (K^{n-1} + K^{n-2} + \dots + 1)\delta$ , and we can see that the true orbit  $\{x_0, f(x_0), \dots, f^N(x_0)\}$  shadows the pseudo-orbit within  $\delta/(1 - K)$ .

Although this hyperbolic map is not sensitive to initial conditions, it is an instructive example. Consider next a diffeomorphism which expands distances, so that  $|f^n(x) - f^n(y)| \geq K^n|x - y|$  for  $K > 1$ . This map is sensitive to initial conditions, yet any pseudo-orbit  $\{x_0, x_1, \dots, x_N\}$  can easily be shadowed. The inverse of the map contracts distances, so the true orbit  $\{f^{-N}(x_N), f^{-N+1}(x_N), \dots, x_N\}$  will shadow the pseudo-orbit within  $\delta/(1 - 1/K)$ .

A general hyperbolic dynamical system is a combination of the above two examples. At each point, some directions are expanding and the rest are contracting. To construct a true orbit, one needs to use information from the beginning of the pseudo-orbit in the contracting directions and from the end of the pseudo-orbit in the expanding directions. This idea is the basis of theorem 3.3.

On the other hand, not every pseudo-orbit can be shadowed. This is not a failure of any particular shadowing procedure. The simplest examples of nonlinear maps provide cases of pseudo-orbits for which there is no corresponding true orbit nearby. Consider the one-dimensional logistic map  $f(x) = 1 - 2x^2$ , shown in figures 2 and 3. The interval  $I = [-1, 1]$  maps onto itself under  $f$  and so is an invariant set. True orbits which begin in  $I$  remain in  $I$  for all time.



**Figure 2.** A pseudo-orbit of  $f(x) = 1 - 2x^2$  which cannot be shadowed. The initial condition is the dot at the origin. An error of size  $\delta$  is made in computing  $f(0)$ , which causes the orbit to eventually approach  $-\infty$ .

Now consider the  $\delta$ -pseudo-orbit which begins with  $x_0 = 0, x_1 = 1 + \delta$ , and which from then on is computed without error. Then  $x_2 = f(x_1) < -1$ , and the pseudo-orbit

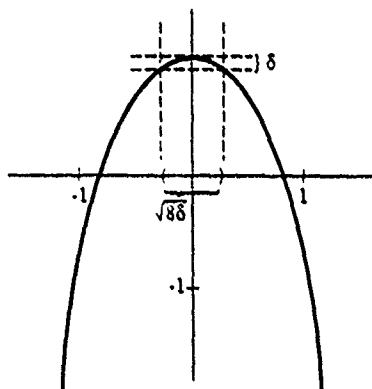


Figure 3. For the map  $f(x) = 1 - 2x^2$ , an initial condition in the open interval of length  $\sqrt{8\delta}$  around zero can be attracted to  $-\infty$  if an error of size  $\delta$  is made.

diverges to  $-\infty$ . See figure 2. Clearly, there is no true orbit of the system  $f$  which shadows the pseudo-orbit by a distance of less than 1. Any true orbit within 1 unit of  $x_0 = 0$  must stay within 1 for all time.

In this simple case, points escape from true behaviour near the critical point, or fold, of the map. Informally, we call such a divergence from legal behaviour a *glitch*. In general, the logistic map  $f(x) = a(1 - x^2) - 1$  will have pseudo-orbits that cannot be shadowed not only for  $a = 2$  as above, but when the parameter  $a$  is less than and within  $\delta$  of 2, where  $\delta$  is the noise level of the process. (This corresponds to the critical value of the fold in figure 2 being between  $1 - \delta$  and 1.) Thus the occurrence of a glitch is a robust phenomenon. The same phenomenon occurs in higher-dimensional chaotic dynamical systems, because of the folds caused by homoclinic tangencies and near-tangencies of stable and unstable manifolds.

How often should we expect glitches? The answer should depend on the noise level  $\delta$ . In the logistic map example  $f(x) = 1 - 2x^2$ , there is an interval of length  $\sqrt{8\delta}$  around zero for which it is possible for an error of size  $\delta$  to cause a glitch. This is illustrated in figure 3. Any initial condition in the designated interval around 0 is susceptible to being mapped to a value greater than 1, and therefore mapped out of  $I$ , towards  $-\infty$ . A computer-generated orbit of that type does not correspond to any true orbit.

If we assume that this interval of length  $\sqrt{8\delta}$  is sampled by the dynamical system, approximately in proportion to its length, we expect a glitch to occur on the order of every  $1/\sqrt{\delta}$  steps. Numerical studies in [6] on two-dimensional maps and the studies of differential equations undertaken for this work roughly support this scaling.

### 3. Shadowing theorem

The theorem can be used to shadow diffeomorphisms or differential equations. To include both cases, we will consider a *dynamical system* to be a sequence  $f_0, \dots, f_N$  of  $C^2$ -diffeomorphisms on  $R^m$  for some positive integer  $N$ .

When attempting to shadow a discrete map  $f$ , we will use  $f_n = f$  for all  $n$ . For a non-autonomous differential equation  $\dot{x} = F(t, x)$ , we would let  $f_n$  be the map on phase space which takes an initial point  $x$  at time  $t$  to the point on the trajectory time at time  $t + h_n$ , where  $h_n$  is the current step size of the ODE solver. If we assume, for simplicity,

that the differential equation is being solved with a constant step size  $h$ , then  $h_n = h$  for all  $n$ . In this case, the ODE solver induces a map called the time- $h$  map of the system.

In the case of an autonomous differential equation, the induced time- $h$  map will be the same for all  $t$ . On the other hand, if the differential equation is non-autonomous, the time- $h$  map will depend on  $t$ . The following definition of an orbit of a dynamical system is made to include both the autonomous and non-autonomous cases.

**Definition 3.1.** Let  $N$  be a positive integer, and let  $f_n : R^m \rightarrow R^m$  be a  $C^2$ -diffeomorphism for each  $0 \leq n < N$ . The finite sequence  $\{y_n\}, n = 0, \dots, N$  of points in  $R^m$  is called an *orbit* of the dynamical system  $\{f_n\}, n = 0, \dots, N-1$  if  $f_n(y_n) = y_{n+1}$  for  $n = 0, \dots, N-1$ . An orbit is sometimes referred to as a *true orbit* to contrast with the notion of pseudo-orbit. The finite sequence  $\{x_n\}$  is called a  $\delta$ -pseudo-orbit of  $\{f_n\}$  if  $|f_n(x_n) - x_{n+1}| < \delta$  for  $n = 0, \dots, N-1$ . The  $\delta$ -pseudo-orbit  $\{x_n\}$  is  $\epsilon$ -shadowed by the orbit  $\{y_n\}$  of the dynamical system  $\{f_n\}$  if  $|x_n - y_n| < \epsilon$  for  $n = 0, \dots, N$ .

Here, as below, we use the Euclidean norm:

$$|v| = \left( \sum_{i=1}^m v_i^2 \right)^{1/2}$$

for a vector  $v = (v_1, \dots, v_m)$ .

We also need to define the concept of moving frame from the point of view of computer simulation. The moving frames we will require will be numerical approximations  $S_n$  and  $U_n$  to the stable tangent space and the unstable tangent space at  $x_n$ , if they exist, and the next best thing, if they do not.

Let  $N$  and  $k$  be positive integers. For each  $n = 0, \dots, N-1$ , let  $J_n$  be a non-singular  $m \times m$  matrix. For each  $n = 0, \dots, N$  let  $\{v_{n1}, \dots, v_{nk}\}_{n=0}^N$  be a set of  $k$  vectors in  $R^m$ , and define  $A_n$  to be the  $m \times k$  matrix with columns  $v_{n1}, \dots, v_{nk}$ .

**Definition 3.2.** The set  $\{v_{n1}, \dots, v_{nk}\}_{n=0}^N$  is called a  $\delta$ -pseudo-frame for the dynamical system  $\{J_n\}$  if for all  $0 \leq n \leq N$ ,

1. The entries of the  $k \times k$  matrix  $A_n^T A_n - I_k$  are no larger than  $\delta$  in absolute value;
2.  $J_n v_{n1}, \dots, J_n v_{nk}$  are each within  $\delta$  of  $\text{range}(A_{n+1})$ .

Informally, we call property 1 of the definition *almost-orthogonality*, and property 2 *consistency*.

The usefulness of this definition for computer-assisted proofs lies in the fact that a  $\delta$ -pseudo-frame consisting of machine-representable numbers can be constructed using standard computational procedures. Assume that we begin with a set of  $k$  vectors  $\{v_{01}, \dots, v_{0k}\}$  in  $R^m$  which form an orthonormal set. (That is, the vectors in the set are mutually orthogonal unit vectors.) Assume further that the components of the vectors  $v_{01}, \dots, v_{0k}$  are machine-representable floating point numbers. Then we use the Gram-Schmidt orthogonalization procedure on the set  $\{J_0(v_{01}), \dots, J_0(v_{0k})\}$ , and define  $\{v_{11}, \dots, v_{1k}\}$  to be the machine-stored vectors that result from this finite-precision computation. (In some cases we found that a more stable form of orthogonalization [4] improved this step.) Continuing in this way for  $0 \leq n \leq N$  we define a  $\delta$ -pseudo-frame for a small number  $\delta$ , such that each vector  $v_{ni}$  in the frame is machine representable.

We can now describe the main theorem. For each  $0 \leq n < N$ , let  $f_n : R^m \rightarrow R^m$  be a  $C^2$ -diffeomorphism. Let  $\{x_n\}_{n=0}^N$  be a  $\delta$ -pseudo-orbit of the dynamical system  $\{f_n\}$ . Define  $J_n = Df(x_n)$  to be the matrix of first partial derivatives of  $f_n$ . Let  $B_1$  (respectively  $B_2$ ) be an upper bound for the absolute values of the first (respectively, second) partial derivatives of the component functions of  $f_n$  and  $f_n^{-1}$  on the union of balls of radius  $\delta^{1/2}$  centred at  $x_n$  for  $n = 0, \dots, N$ . Set  $B = \max\{2, B_1, B_2\}$ . For positive integers  $k+l=m$ , let  $\{s_{n1}, \dots, s_{nk}\}_{n=0}^N$  (respectively,  $\{u_{n1}, \dots, u_{nl}\}_{n=0}^N$ ) be a  $\delta$ -pseudo-frame for  $\{J_n^{-1}\}$  (respectively,  $\{J_n\}$ ) such that  $\{s_{n1}, \dots, s_{nk}, u_{n1}, \dots, u_{nl}\}$  spans  $R^m$  for each  $n$ .

Define the subspaces  $S_n = \text{span}\{s_{n1}, \dots, s_{nk}\}$ ,  $U_n = \text{span}\{u_{n1}, \dots, u_{nl}\}$ , and define  $\theta_n$  to be the angle between  $S_n$  and  $U_n$ . Let  $r_n$  and  $t_n$  be numbers satisfying

$$\begin{aligned} |J_n y| &\leq r_n |y| && \text{for } y \in S_n \\ |J_n^{-1} z| &\leq t_n |z| && \text{for } z \in U_{n+1}. \end{aligned}$$

Define  $C_0 = 0, C_n = \csc \theta_n + r_{n-1} C_{n-1}$  for  $n > 0$ . Similarly, define  $D_N = 0, D_n = \csc \theta_n + t_n D_{n+1}$  for  $n < N$ .

**Theorem 3.3.** Let  $\{x_n\}_{n=0}^N$  be a  $\delta$ -pseudo-orbit for the dynamical system  $\{f_n\}$  on  $R^m, m \geq 2$ , and assume that  $\delta < 1/20m^2$ . If

$$\max\{C_n, D_n\} \leq \frac{1}{m^{5/2} B^2 \sqrt{\delta}}$$

for all  $n = 0, \dots, N$ , then there exists an orbit  $\{w_n\}$  of  $\{f_n\}$  such that  $|x_n - w_n| < \sqrt{\delta}$  for  $n = 0, \dots, N$ .

Theorem 3.3 gives an alternative approach to Bowen's shadowing lemma [2]. Let  $f : R^m \rightarrow R^m$  be a  $C^2$ -diffeomorphism. A compact invariant set  $\Lambda$  is called *hyperbolic* if there is a continuous splitting of the tangent space  $T_x R^m = E_x^s \oplus E_x^u$  for  $x \in \Lambda$ , and positive constants  $\lambda < 1, C > 0$  such that

1.  $Df(x)(E_x^s) = E_{f(x)}^s$
2.  $Df(x)(E_x^u) = E_{f(x)}^u$
3.  $|Df^n(x)(v)| \leq C\lambda^{-n}|v|$  for  $v \in E_x^s$
4.  $|Df^{-n}(x)(v)| \leq C\lambda^{-n}|v|$  for  $v \in E_x^u$

for all  $x \in \Lambda$  and for all  $n \geq 0$ .

**Theorem 3.4.** [2]. Assume  $\Lambda$  is a hyperbolic set for  $f$ . For every  $\varepsilon > 0$  there is a  $\delta > 0$  so that every  $\delta$ -pseudo-orbit in  $\Lambda$  can be  $\varepsilon$ -shadowed.

Theorem 3.4 is a direct consequence of theorem 3.3 (see [9]).

#### 4. Computer-assisted shadowing

In this section we describe a computer algorithm which uses the above theorem 3.3 to verify the existence of true orbits of a dynamical system near the pseudo-orbit determined by a numerical computation. Along with the pseudo-orbit being computed, there are some auxiliary calculations to be made to check that the hypotheses of the theorem are satisfied. Namely, it is necessary to find upper bounds for the constants  $B, \csc \theta, r_n, t_n$ , and finally  $C_n$  and  $D_n$ . We next describe these auxiliary calculations, which if successful provide a computer-assisted proof of the existence of a true orbit.

#### 4.1. Construction of stable and unstable frames

The algorithm works best when the  $\delta$ -pseudo-frames  $\{s_{n1}, \dots, s_{nk}\}_{n=0}^N$  and  $\{u_{n1}, \dots, u_{nl}\}_{n=0}^N$  are chosen to approximately encompass the stable and unstable directions, respectively, for the dynamical system  $\{f_n\}$  at the particular map  $f_n$ . One way to accomplish this is as follows. Begin with an orthonormal set  $\{u_{01}, \dots, u_{0l}\}$  of vectors in  $R^m$  chosen arbitrarily. Inductively define the orthonormal set  $\{u_{n+1,1}, \dots, u_{n+1,l}\}$  to be the computed results of applying the technique of Gram-Schmidt orthogonalization, followed by normalization, to the set  $\{Df_n(x_n)u_{n1}, \dots, Df_n(x_n)u_{nl}\}$ . Because of computer round-off, these computations will be only approximate, which is not important.

For which  $\delta$  is  $\{u_{n1}, \dots, u_{nl}\}_{n=0}^N$  a  $\delta$ -pseudo-frame? It is straightforward to find a  $\delta$  for which both parts of definition 3.2 are satisfied. Part 1 is easily checked with the computed  $\{u_{n1}, \dots, u_{nl}\}_{n=0}^N$  and depends on the residual error of the Gram-Schmidt orthogonalization. In most cases of following a trajectory of a system of ordinary differential equations, the  $\delta$  will be determined by part 2 of definition 3.2, which depends on the error bound of the ODE solver being used to follow the tangent vectors along the pseudo-orbit.

The frame  $\{s_{n1}, \dots, s_{nk}\}_{n=0}^N$  is defined analogously. Begin with an arbitrary orthonormal set  $\{s_{01}, \dots, s_{0k}\}$  from  $R^m$ . Given  $\{s_{n1}, \dots, s_{nk}\}$  for  $n \leq N$ , apply Gram-Schmidt to the set  $\{Df_{n-1}^{-1}(x_{n-1})s_{n1}, \dots, Df_{n-1}^{-1}(x_{n-1})s_{nk}\}$ . The stored values of the resulting computation are  $\{s_{n-1,1}, \dots, s_{n-1,k}\}$ , by definition.

The calculation of  $\csc \theta_n$ , where  $\theta_n$  is the angle between  $S_n$  and  $U_n$ , is simple if the dimension  $m$  is small, but for higher dimensions the following scheme may be helpful. Define  $A_n$  to be the  $m \times m$  matrix whose columns are  $\{s_{n1}, \dots, s_{nk}, u_{n1}, \dots, u_{nl}\}$ , and let  $B_n = A_n^{-1}$ . Let  $B_n^s$  be the  $m \times m$  matrix whose top  $k$  rows are the same as those of  $B_n$  and whose bottom  $l$  rows are filled with zeros. Let  $B_n^u$  be the  $m \times m$  matrix whose top  $k$  rows are filled with zeros and whose bottom  $l$  rows are the same as those of  $B_n$ . Note that  $B_n = B_n^s + B_n^u$ .

Now define  $S_n = A_n B_n^s$  and  $U_n = A_n B_n^u$ . It is clear that  $S_n$  and  $U_n$  are projections onto  $S_n$  and  $U_n$ , respectively, and that

$$S_n + U_n = A_n (B_n^s + B_n^u) = I.$$

Further,  $S_n$  and  $U_n$  are the unique  $m \times m$  matrices with these properties.

It is a standard fact that  $\csc \theta_n = |S_n| = |U_n|$ , where as usual we use the Euclidean matrix norm. This scheme provides a computationally stable method for computing a strict upper bound on  $\csc \theta_n$ , which is necessary for bounding the  $C_n$  and  $D_n$ .

#### 4.2. Calculation of $r_n$ and $t_n$

We have defined  $r_n$  to be a positive number that bounds the growth of  $f_n$  in the direction  $S_n$  at  $x_n$ . That is,  $r_n$  satisfies  $|J_n y| \leq r_n |y|$  for vectors  $y$  in  $S_n$ . Such a number is impossible to find by measuring  $|J_n y|$  on a general basis of  $S_n$ . This is the reason that almost-orthogonal frames are needed in the calculation. Lemma 4.1, using  $A = J_n$  and  $W = S_n$ , shows how to find an upper bound on  $r_n$  solely using information about the action of  $J_n$  on the almost-orthogonal basis of  $S_n$ . Analogously,  $t_n$  can be found by using lemma 4.1 with  $A = J_n^{-1}$ , on the subspace  $U_{n+1}$ .

**Lemma 4.1.** Let  $A$  be an  $m \times m$  matrix and  $W$  a subspace of  $R^m$  with basis  $\{w_1, \dots, w_k\}$ . Let  $W$  be the  $m \times k$  matrix with columns  $\{w_1, \dots, w_k\}$ . Then

$$\max_{v \in W, \|v\|=1} |Av| \leq \frac{|AW|}{\sqrt{1 - |W^T W - I|}}$$

when the right-hand side exists.

Lemma 4.1 is proved in section 5.

#### 4.3. Calculation of $C_n$ and $D_n$

Computing  $C_n$  and  $D_n$  appears simple once  $\csc \theta_n$ ,  $r_n$  and  $t_n$  are known. There are two more details, however, that greatly reduce the data requirements of this task. In applications of this algorithm, it is typical for  $N$ , the number of points in the pseudo-orbit, to be of the order of several million. On the other hand, we have previously suggested that the computation of the stable frame  $\{s_{n1}, \dots, s_{nk}\}_{n=0}^N$  (and therefore  $\theta_n$ ) be done by beginning with a random frame at  $n = N$ , and applying  $J_n^{-1}$  to create frames  $N-1, \dots, 0$ . To avoid the problem of storing all frames simultaneously, we suggest building  $\{s_{n1}, \dots, s_{nk}\}_{n=0}^N$  in pieces of length  $N_1 < N$ . For example, we found  $N_1 = 5000$  to be reasonable.

The idea is to find each block of 5000 nearly-orthogonal bases by stopping after each block of 5000 points in the pseudo-orbit, finding the next 1000 points, and then applying  $J_n^{-1}$  6000 times to a random starting orthogonal basis to produce stable directions, and then go on to the next block of 5000. In all cases we have tried, the stable frame produced this way satisfied definition 3.2 within the prescribed  $\delta$ .

The second problem is deciding whether the recurrence relation  $D_n$  stays within the bound of the theorem, given that it is defined beginning at the end of the trajectory. The following simple lemma shows how to verify the bound on  $D_n$  in forward time. In short, a new recurrence relation  $E_n$  is defined which is computed in forward time. The lemma shows how to tell by computing  $E_n$  whether  $D_n$  violates a given bound.

**Lemma 4.2.** Let  $D_N = 0$ ;  $D_n = a_n + b_n D_{n+1}$  be a recurrence relation for  $n = 0, \dots, N$  and let  $A$  be a real number. Define another recurrence relation  $E_0 = A$ ;  $E_{n+1} = \min\{(E_n - a_n)/b_n, A\}$  for  $n = 0, \dots, N$ . If  $E_n \geq 0$  for  $n = 0, \dots, N$ , then  $D_n \leq A$  for  $n = 0, \dots, N$ .

#### 4.4. Calculation of $B$

The calculation of  $B$ , the upper bound on the magnitudes of the first and second partial derivatives of the  $f_n$ , is normally trivial if we are given the explicit map. In more interesting cases, we are following the (possibly time-dependent) flow of a system of differential equations, and need bounds on the derivatives of the time- $h$  map for step size  $h$ . It is this map which is being approximated by the numerical ODE solver.

To this end, consider the first-order system

$$\dot{y} = F(t, y)$$

where  $y$  is a vector in  $R^m$  and  $t$  denotes the independent variable. Define  $g(t, s, z)$  to be the value of the solution with initial condition  $y(s) = z$  at time  $t$ . Then the time- $h$  map of the differential equation which maps the value at time  $t_0$  to the value at time  $t_0 + h$  is given by

$$f_{t_0, h}(y) = g(t_0 + h, t_0, y).$$

The following lemma establishes upper bounds on the partial derivatives of the  $m$  component functions of  $f = f_{t_0, h} = (f_1, \dots, f_m)$ .

*Lemma 4.3.*

1. Define  $E_1 = \max_{i,j} \left| \frac{\partial F_i}{\partial y_j} \right|$ . Then  $\max_{i,j} \left| \frac{\partial f_i}{\partial y_j} \right| \leq e^{hmE_1}$ .
2. Define  $E_2 = \max_{i,j,k} \left| \frac{\partial^2 F_i}{\partial y_j \partial y_k} \right|$ . Then  $\max_{i,j,k} \left| \frac{\partial^2 f_i}{\partial y_j \partial y_k} \right| \leq hm^2 E_2 e^{3hmE_1}$ .

The proof is an exercise in using the Gronwall inequality (see for example lemma 4.1 of [8]) on the first and second variational equations of the system.

#### 4.5. Quadrature method

To apply theorem 3.3 to a differential equation such as the forced damped pendulum, we need a quadrature method which has high accuracy, and which has an explicit error formula. The former is necessary to allow application of the theorem with a reasonably small  $\delta$  (and therefore a long shadowing time). The latter is necessary to assure that we have a rigorous bound on  $\delta$ .

The simplest method that satisfies these two criteria is the Taylor method. The formula for the one-step error is explicit, being essentially the Taylor remainder. However, the major difficulty with implementation of the Taylor methods in general is that they require explicit differentiation of the right-hand side of the differential equation. Thus, applying the seventh-order Taylor method to the differential equation

$$\ddot{y} + a\dot{y} + \sin y = b \cos t \quad (1)$$

evidently requires differentiating the differential equation five times. The formulae fill a few pages.

Fortunately, there is a trick which allows application of the Taylor method as an ODE solver without doing the symbolic calculation of higher derivatives of the differential equation. We illustrate the trick in terms of equation (1). Set  $z_1 = \sin y$  and  $z_2 = \cos y$ . Then

$$\begin{aligned} \dot{z}_1 &= (\cos y)\dot{y} = z_2\dot{y} \\ \dot{z}_2 &= (-\sin y)\dot{y} = -z_1\dot{y}. \end{aligned}$$

Now given a point  $(y, \dot{y})$  in phase space at time  $t$ , we show how to calculate the higher derivatives of  $y$  at time  $t$ . First of all, we can calculate  $z_1, z_2$  from the definitions and  $\dot{y}$  from equation (1). Then, for  $i \geq 1$ , we recursively calculate

$$\begin{aligned} z_1^{(i)} &= (z_2\dot{y})^{(i-1)} = \sum_{j=0}^{i-1} \binom{i-1}{j} z_2^{(j)} y^{(i-j)} \\ z_2^{(i)} &= (-z_1\dot{y})^{(i-1)} = -\sum_{j=0}^{i-1} \binom{i-1}{j} z_1^{(j)} y^{(i-j)} \\ y^{(i+2)} &= -ay^{(i+1)} - z_1^{(i)} + b(\cos t)^{(i)} \end{aligned}$$

using the differential equation and the product rule of Leibniz. The higher derivatives of  $y$  at time  $t$  are therefore known, so we can apply the Taylor method of arbitrary order with no symbolic calculation beforehand. A similar trick applies to the variational equation of (1). We applied the seventh-order Taylor method to follow solutions of both the differential equation and the variational equation. The latter is necessary for calculating a rigorous  $\delta$ -pseudo-frame for the computer-generated trajectory.

## 5. Examples

As a first example, consider the Hénon map

$$f(x, y) = (a - x^2 + by, x)$$

of the plane. For parameter values  $a = 1.4$ ,  $b = -0.3$ , this map has an apparently chaotic orbit. Using the method described above, a computer-generated  $\delta$ -pseudo-orbit with initial condition  $(0, 0)$  and  $\delta = 10^{-14}$  was found to have a true orbit within  $10^{-7}$  for over one million iterates. Similar statements apply for other initial conditions, and for other parameter values.

The pseudo-orbits generated by our computer satisfied  $|x| \leq 2$ ,  $|y| \leq 2$  in every case. In this range, the magnitudes of the first partial derivatives of  $f = (f_1, f_2)$  and the easily-computed inverse  $f^{-1} = (g_1, g_2)$  are bounded above by 4. The magnitudes of the second partial derivatives are bounded by 2. Therefore we used  $m = 2$ ,  $B = 4$  in the hypotheses of theorem 3.3.

This map was originally shadowed in [6], and similar results were reported. In that paper, a different approach was taken, which uses 96-bit arithmetic (machine-epsilon =  $10^{-28}$ ) to verify shadowing of a  $\delta$ -pseudo-orbit calculated in 48-bit arithmetic, i.e. with  $\delta = 10^{-14}$ . The method of the present paper does not require such higher precision for this map.

This point becomes especially relevant when systems are studied that are inherently more difficult to shadow. Consider the forced damped pendulum, which satisfies the differential equation

$$\ddot{y} + a\dot{y} + \sin y = b \cos t.$$

To achieve good shadowing results for this differential equation we needed to generate a  $\delta$ -pseudo-trajectory with  $\delta = 10^{-18}$ . We accomplish this by using a seventh-order one-step quadrature method with an explicit truncation error formula, using a step size of  $h = \pi/1000$ . The implementation details of the quadrature method are given in section 4.5. The fact that the quadrature error formula is explicit is critical. Without it we could not get a rigorous bound on  $\delta$ .

For the forced damped pendulum with parameters  $a = 0.2$  and  $b = 2.4$ , there is an apparently chaotic trajectory with initial conditions  $y(0) = \dot{y}(0) = 0$ . Using theorem 3.3, we proved the existence of a true trajectory within  $10^{-9}$  of the computer-generated trajectory for time  $t$  ranging from 0 to  $10^4\pi$ . This trajectory corresponds to  $10^7$  time steps of the ODE solver. Again, there are similar results for other initial conditions, and other values of  $a$  and  $b$ .

The maps  $f_n$  used in theorem 3.3 were the time- $h$  maps of the non-autonomous differential equation, where  $h = \pi/1000$ . The derivation of  $B$  for the forced damped pendulum uses lemma 4.3. Write the pendulum equation as a first-order system. Then in lemma 4.3 is

$$F(t, y_1, y_2) = (y_2, -\sin y_1 - ay_2 + b \cos t).$$

It is easy to check that the first and second partial derivatives of  $F$  with respect to  $y_1$  and  $y_2$  are bounded in absolute value by 1, so that  $E_1 = E_2 = 1$ . Lemma 4.3 says that  $B = \max\{2, e^{2h}, 4he^{6h}\}$ . Since  $h = 10^{-18}$ , we use  $B = 2$ ,  $m = 2$  in the hypotheses of theorem 3.3. Note that the inverse of a time- $h$  map is a time- $-h$  map, so that the same  $B$  works for  $f_n^{-1}$ .

## 6. Proof of theorem

The convention in this section, as in the entire paper, is that all vector and matrix norms are  $l^2$  (Euclidean) norms. The norm of an  $m \times m$  matrix  $A$  is defined in terms of the vector norm, as follows:

$$|A| = \max_{v \in \mathbb{R}^m, |v|=1} |Av|.$$

It follows from the definitions that  $|A| = \sqrt{\sigma(A^T A)}$ , where  $\sigma(B)$  denotes the maximum absolute value of the eigenvalues of the symmetric matrix  $B$ .

**Lemma 6.1.** If  $A$  is an  $m \times m$  matrix whose entries are at most  $\delta$  in absolute value, then  $|A| \leq m\delta$ .

*Proof.*  $|A| \leq |A|_F$ , where  $|A|_F^2 = \sum_{i=1}^m \sum_{j=1}^m a_{ij}^2$ . See [4].

**Lemma 6.2.** If  $W$  is an  $m \times k$  matrix and  $x = Wy$ , then

$$|y| \leq \frac{|x|}{\sqrt{1 - |W^T W - I|}}$$

when the right-hand side exists.

*Proof.*

$$\begin{aligned} |y|^2 &= y^T y - y^T W^T W y + |x|^2 \\ &= y^T (I - W^T W) y + |x|^2 \\ &\leq |y| |(I - W^T W) y| + |x|^2 \\ &\leq |I - W^T W| |y|^2 + |x|^2. \end{aligned}$$

*Proof of lemma 4.1.* Let  $x \in \mathbf{W}$ . Then  $x = Wy$ , and  $Ax = AWy$ . By lemma 6.2,

$$\max_{|x|=1, x \in \mathbf{W}} |Ax|^2 = \frac{1}{1 - |W^T W - I|} \max_{|y| \leq 1, y \in \mathbb{R}^k} |AWy|^2$$

and

$$\max_{y \in \mathbb{R}^k, |y|=1} |AWy|^2 = \max_{y \in \mathbb{R}^k, |y|=1} y^T (AW)^T AW y = \sigma((AW)^T AW).$$

**Lemma 6.3.** Let  $\{v_{n1}, \dots, v_{nk}\}_{n=0}^1$  be a  $\delta$ -pseudo-frame for the matrix  $J$ , where  $\delta < 3/(4k)$ . Let  $A_n$  be the matrix with columns  $\{v_{n1}, \dots, v_{nk}\}$ . For each  $v \in \text{range } A_0$  there is a  $w \in \text{range } A_1$  such that  $|Jv - w| \leq 2\sqrt{k}\delta|v|$ .

*Proof.* Let  $v = \sum_{i=1}^k c_i v_{0i}$ ; that is,  $v = A_0 c$ . Define  $w = A_1 c = \sum_{i=1}^k c_i v_{1i}$ . Then

$$\begin{aligned} |Jv - w| &= \left| \sum_{i=1}^k c_i (Jv_{0i} - v_{1i}) \right| \\ &\leq \delta \sum_{i=1}^k |c_i| \leq \delta \sqrt{k} |c| \\ &\leq \frac{\delta \sqrt{k} |v|}{\sqrt{1 - |A_0^T A_0 - I|}} \leq 2\delta \sqrt{k} |v| \end{aligned}$$

where the last line follows from lemma 6.2. □

The next two lemmas refer to a  $C^2$ -map  $f$  which maps a convex subset  $S$  of  $R^m$  to  $R^m$ . Define  $B_1$  (respectively,  $B_2$ ) to be an upper bound on the magnitude of all first (respectively, second) partial derivatives of all component functions of  $f$  on  $S$ . Assume that  $x$  and  $x+h$  lie in  $S$ .

*Lemma 6.4.*

1.  $|f(x+h) - f(x)| \leq mB_1|h|$ .
2.  $|Df(x+h) - Df(x)| \leq m\sqrt{m}B_2|h|$ .

*Proof.* For a scalar function  $g$ ,

$$|g(x+h) - g(x)| \leq \max \left| \frac{\partial g}{\partial x_j} \right| \sum_{j=1}^m |h_j| \leq |h| \sqrt{m} \max \left| \frac{\partial g}{\partial x_j} \right|.$$

Applying this to each entry of the vector  $f$ , and the matrix  $Df$ , respectively, one gets the stated estimates.

*Lemma 6.5.*

$$|f(x+h) - f(x) - Df(x)h| \leq \frac{m\sqrt{m}B_2|h|^2}{2}.$$

*Proof.* Each component  $g$  of  $f$  satisfies

$$|g(x+h) - g(x) - Df(x)h| \leq \frac{m|h|^2 B_2}{2}$$

from which the result follows easily.  $\square$

Now assume that  $\{s_{n1}, \dots, s_{nk}\}_{n=0}^N$  and  $\{u_{n1}, \dots, u_{nl}\}_{n=0}^N$  are  $\delta$ -pseudo-frames for the dynamical system  $\{J_n\}$  on  $R^m$ , where  $k+l=m$ . Let  $B$  be an upper bound for the magnitude of all entries of the  $J_n$ . Let  $S_n$  and  $U_n$  be the subspaces spanned by the moving frames and let  $S_n$  and  $U_n$  be the projections onto the subspaces such that  $S_n + U_n = I$ .

*Lemma 6.6.*

1. For  $u \in U_n$ ,
  - (a)  $|S_{n+1}J_n u| \leq 2m^{1/2}\delta|S_{n+1}||u|$ , and
  - (b)  $|S_{n-1}J_{n-1}^{-1}u| \leq 2m^{3/2}B\delta|S_{n-1}||u|$ .
2. For  $s \in S_n$ ,
  - (a)  $|U_{n-1}J_{n-1}^{-1}s| \leq 2m^{1/2}\delta|U_{n+1}||s|$ , and
  - (b)  $|U_{n+1}J_n s| \leq 2m^{3/2}B\delta|U_{n+1}||s|$ .

*Proof.* To prove 1, we use the fact that  $\{u_{n1}, \dots, u_{nl}\}_{n=0}^N$  is a  $\delta$ -pseudo-frame for  $\{J_n\}_{n=0}^{N-1}$ . If we are given  $u \in U_n$ , there is  $w \in U_{n+1}$  such that  $|J_n u - w| \leq 2\sqrt{l}\delta|u|$ , by lemma 6.3.

$$\begin{aligned} |S_{n+1}J_n u| &= |S_{n+1}J_n u - S_{n+1}w + S_{n+1}w| \\ &= |S_{n+1}J_n u - S_{n+1}w| \\ &\leq |S_{n+1}||J_n u - w| \\ &\leq 2\sqrt{m}\delta|S_{n+1}||u|. \end{aligned}$$

Secondly, we use the fact that  $\{u_{n1}, \dots, u_{nl}\}_{n=0}^N$  is a  $\delta mB$ -pseudo-frame for  $\{J_n^{-1}\}_{n=0}^{N-1}$ . Given  $u \in U_n$ , there is  $w \in U_{n-1}$  such that  $|J_{n-1}^{-1}u - w| \leq 2\sqrt{l}mB\delta|u|$ .

$$\begin{aligned} |S_{n-1}J_{n-1}^{-1}u| &= |S_{n-1}J_{n-1}^{-1}u - S_{n-1}w + S_{n-1}w| \\ &= |S_{n-1}J_{n-1}^{-1}u - S_{n-1}w| \\ &\leq |S_{n-1}||J_{n-1}^{-1}u - w| \\ &\leq 2m^{3/2}B\delta|S_{n-1}||u|. \end{aligned}$$

Part 2 is similar. □

*Proof of theorem 3.3.* For each  $0 \leq n < N$ ,  $f_n : R^m \rightarrow R^m$  is a  $C^2$ -diffeomorphism. Let  $\{x_n\}_{n=0}^N$  be a  $\delta$ -pseudo-orbit of the dynamical system  $\{f_n\}$ . Define  $J_n = Df(x_n)$ . Let  $B_1$  (respectively,  $B_2$ ) be an upper bound for the absolute value of the first (respectively, second) partial derivatives of the component functions of  $f_n$  and  $f_n^{-1}$  on the union of balls of radius  $\delta^{1/2}$  centred at  $x_n$  for  $n = 0, \dots, N$ . Set  $B = \max\{2, B_1, B_2\}$ . For positive integers  $k+l = m$ , let  $\{s_{n1}, \dots, s_{nk}\}_{n=0}^N$  (respectively,  $\{u_{n1}, \dots, u_{nl}\}_{n=0}^N$ ) be a  $\delta$ -pseudo-frame for  $\{J_n^{-1}\}$  (respectively,  $\{J_n\}$ ) such that  $\{s_{n1}, \dots, s_{nk}, u_{n1}, \dots, u_{nl}\}$  spans  $R^m$  for each  $n$ .

Define  $S_n = \text{span}\{s_{n1}, \dots, s_{nk}\}$ ,  $U_n = \text{span}\{u_{n1}, \dots, u_{nl}\}$ , and define  $\theta_n$  to be the angle between  $S_n$  and  $U_n$ . Let  $S_n$  (respectively,  $U_n$ ) be the (unique) projection onto  $S_n$  (respectively,  $U_n$ ) such that  $S_n + U_n = I$ . Recall that  $|S_n| = |U_n| = \csc \theta_n$ . Let  $r_n$  and  $t_n$  be numbers satisfying

$$\begin{aligned} |J_n y| &\leq r_n |y| & \text{for } y \in S_n \\ |J_n^{-1} z| &\leq t_n |z| & \text{for } z \in U_{n+1}. \end{aligned}$$

Define  $x_n^0 = x_n$ ,  $y_0^i = z_0^i = 0$  for  $i = 0, 1, 2, \dots$ , and define

$$y_n^i = S_n(f(x_{n-1}^i) - x_n^i + J_{n-1}y_{n-1}^i) \quad \text{for } n = 1, \dots, N. \quad (2)$$

$$z_n^i = U_n(f^{-1}(x_{n+1}^i) - x_n^i + J_n^{-1}z_{n+1}^i) \quad \text{for } n = 0, \dots, N-1. \quad (3)$$

$$x_n^{i+1} = x_n^i + y_n^i + z_n^i \quad \text{for } n = 0, \dots, N. \quad (4)$$

The sequence  $\{x_n^i\}_{n=0}^N$  is the result of applying the refinement technique  $i$  times to the original pseudo-orbit  $\{x_n\}_{n=0}^N$ . Define  $p$  by  $\delta^{-p} = m^{5/2}B^2$ . Let  $C_0 = 0$ ,  $C_n = |S_n| + r_{n-1}C_{n-1}$  for  $n > 0$ . Similarly, let  $D_0 = 0$ ,  $D_n = |U_n| + t_n D_{n+1}$  for  $n < N$ .

**Lemma 6.7.** Assume that  $C_n \leq \delta^{p-1/2}$  and  $D_n \leq \delta^{p-1/2}$  for  $n = 0, \dots, N$ . Then for  $n = 0, \dots, N$  and  $i \geq 0$ :

- (a)  $|y_n^i| \leq 2^{-i} \delta C_n \leq 2^{-i} \delta^{p+1/2}$
- (b)  $|z_n^i| \leq 2^{-i} \delta D_n \leq 2^{-i} \delta^{p+1/2}$
- (c)  $|x_n^{i+1} - x_n^i| \leq 2^{1-i} \delta^{p+1/2}$
- (d)  $|x_n^{i+1} - x_n^0| \leq 4\delta^{p+1/2} \leq \frac{4\sqrt{\delta}}{m^{5/2}B^2} \leq \sqrt{\delta}$

**Proof.** Statement (d) follows from (c). Statements (a), (b) and (c) are proved by double induction on  $i$  and  $n$ . If  $i = 0$ :

- (a)  $|y_0^0| = 0$ , and for  $n > 0$ ,

$$|y_n^0| \leq |S_n| \delta + r_{n-1} |y_{n-1}^0| \leq |S_n| \delta + r_{n-1} \delta C_{n-1} \leq C_n \delta.$$

- (b)  $|z_n^0| \leq D_n \delta$ , by reasoning similar to (a).
- (c)  $|x_n^1 - x_n^0| \leq |y_n^0| + |z_n^0| \leq \delta(C_n + D_n) \leq \delta 2\delta^{p-1/2} = 2\delta^{p+1/2}$ .

Now we assume that (a) holds for  $i-1$ , and prove it for case  $i$ . We induct on  $n$ . The  $n=0$  case is trivial, since  $|y_0^i| = 0$ . Assume that (a) holds for the case  $i, n-1$  and prove it for  $i, n$ :

$$\begin{aligned} y_n^i &= S_n(f_{n-1}(x_{n-1}^i) - x_n^i + J_{n-1}y_{n-1}^i) \\ &= S_n(f_{n-1}(x_{n-1}^i) - f_{n-1}(x_{n-1}^{i-1}) + f_{n-1}(x_{n-1}^{i-1}) \\ &\quad - x_n^{i-1} - y_n^{i-1} - z_n^{i-1} + J_{n-1}y_{n-1}^i) \\ &= S_n[f_{n-1}(x_{n-1}^i) - f_{n-1}(x_{n-1}^{i-1}) - Df_{n-1}(x_{n-1}^{i-1})(x_{n-1}^i - x_{n-1}^{i-1}) \\ &\quad + (Df_{n-1}(x_{n-1}^{i-1}) - J_{n-1})(y_{n-1}^{i-1} + z_{n-1}^{i-1}) + J_{n-1}z_{n-1}^{i-1} + J_{n-1}y_{n-1}^i] \end{aligned}$$

where we have used the facts that  $S_n(z_n^{i-1}) = 0$  and

$$S_n(x_n^{i-1} + y_n^{i-1} - f_{n-1}(x_{n-1}^{i-1})) = S_n(J_{n-1}y_{n-1}^{i-1})$$

by the definition of  $y_n^{i-1}$ .

We will bound the Euclidean norm of each of the four terms of the last sum separately.

1.

$$\begin{aligned} |S_n(f_{n-1}(x_{n-1}^i) - f_{n-1}(x_{n-1}^{i-1}) - Df_{n-1}(x_{n-1}^{i-1})(x_{n-1}^i - x_{n-1}^{i-1}))| &\leq |S_n| \frac{m^{3/2}B_2}{2} |x_{n-1}^i - x_{n-1}^{i-1}|^2 \\ &\leq |S_n| \frac{m^{3/2}B_2}{2} (2^{2-i} \delta^{p+1/2})^2 \\ &\leq |S_n| \delta 2^{-i-2} 32B_2 2^{-i} m^{3/2} \delta^{2p} \\ &\leq |S_n| \delta 2^{-i-2} \end{aligned}$$

since  $m \geq 2$ ,  $B \geq 2$  implies that  $\delta^{-2p} > m^5 B^4 = 32Bm^{3/2} (m^{7/2} B^3)/32 > 32Bm^{3/2}$ .

2.

$$\begin{aligned}
|S_n(Df_{n-1}(x_{n-1}^{i-1}) - J_{n-1})(y_{n-1}^{i-1} + z_{n-1}^{i-1})| &\leq |S_n| m^{3/2} B_2 |x_{n-1}^{i-1} - x_{n-1}| |y_{n-1}^{i-1} + z_{n-1}^{i-1}| \\
&\leq |S_n| m^{3/2} B_2 4\delta^{p+1/2} 2^{2-i} \delta^{p+1/2} \\
&\leq |S_n| \delta 2^{-i-2} 32\delta B_2 \delta^{2p} m^{3/2} \\
&\leq |S_n| \delta 2^{-i-2}
\end{aligned}$$

since  $\delta^{-2p} > 32B_2 m^{3/2}$ .

3.

$$\begin{aligned}
|S_n J_{n-1} z_{n-1}^{i-1}| &\leq 2|S_n| \sqrt{m} \delta |z_{n-1}^{i-1}| \\
&\leq 2\delta^{p-1/2} \sqrt{m} \delta 2^{-i+1} \delta^{p+1/2} \\
&= 16\delta^{2p} \sqrt{m} 2^{-i-2} \delta \\
&\leq 2^{-i-2} \delta
\end{aligned}$$

since  $\delta^{-2p} > 16\sqrt{m}$ .

4.

$$\begin{aligned}
|S_n J_{n-1} y_{n-1}^i| &\leq |J_{n-1} y_{n-1}^i| + |U_n J_{n-1} y_{n-1}^i| \\
&\leq r_{n-1} |y_{n-1}^i| + 2m^{3/2} B \delta |U_n| 2^{-i} \delta^{p+1/2} \\
&\leq r_{n-1} |y_{n-1}^i| + 8\delta^{2p} m^{3/2} B 2^{-i-2} \delta \\
&\leq r_{n-1} |y_{n-1}^i| + 2^{-i-2} \delta
\end{aligned}$$

since  $\delta^{-2p} > 8m^{3/2} B$ .

Adding up the four bounds we have

$$\begin{aligned}
|y_n^i| &\leq 4|S_n| \delta 2^{-i-2} + r_{n-1} |y_{n-1}^i| \\
&\leq |S_n| \delta 2^{-i} + r_{n-1} 2^{-i} \delta C_{n-1} \\
&= C_n \delta 2^{-i}.
\end{aligned}$$

This proves (a) for the case  $i$ . The proof of (b) is similar, except that we use descending induction on  $n$ . The  $n = N$  initial case of (b) is trivial, since  $|z_N^i| = 0$ . Finally, (c) is a simple consequence of (a) and (b).  $\square$

Lemma 6.7 shows that for each  $n$ ,  $y_n^i \rightarrow 0$ ,  $z_n^i \rightarrow 0$ , and that  $\{x_n^i\}_{i=0}^\infty$  is a Cauchy sequence. Therefore  $x_n^i$  converges to some  $w_n$  in  $R^m$ . The sequence  $\{w_n\}_{n=0}^N$  is the limit of the refinement process applied to  $\{x_n\}_{n=0}^N$ . Moreover, lemma 6.7 (d) implies that  $|w_n - x_n| < \delta^{1/2}$ .

We will complete this section by showing that  $f_n(w_n) = w_{n+1}$  for  $n = 0, \dots, N-1$ , so that the  $\{w_n\}$  represents a true shadowing orbit of  $\{x_n\}$ . According to equations (2) and (3),

$$S_{n+1}(f_n(w_n) - w_{n+1}) = 0 \quad (5)$$

and

$$U_n(f_n^{-1}(w_{n+1}) - w_n) = 0 \quad (6)$$

for  $n = 0, \dots, N-1$ . Furthermore, we have

$$\begin{aligned} |f_n(w_n) - w_{n+1}| &\leq |f_n(w_n) - f_n(x_n)| + |f_n(x_n) - x_{n+1}| + |x_{n+1} - w_{n+1}| \\ &\leq mB_1|w_n - x_n| + \delta + 4\delta^{p+1/2} \\ &\leq (mB_1 + 1)4\delta^{p+1/2} + \delta \\ &\leq \delta^{1/2} \left( \delta^{1/2} + \frac{4(mB + 1)}{m^{5/2}B^2} \right) \\ &\leq \delta^{1/2} \left( \frac{1}{\sqrt{80}} + \frac{4(4+1)}{2^{5/2}2^2} \right) \\ &< \delta^{1/2}. \end{aligned}$$

A similar calculation shows that

$$|f_n^{-1}(w_{n+1}) - w_n| < \delta^{1/2}.$$

Secondly, corollary to this calculation are the facts that

$$|f_n(w_n) - x_{n+1}| < \delta^{1/2}$$

and

$$|f_n^{-1}(w_{n+1}) - x_n| < \delta^{1/2}$$

Therefore  $f_n(w_n)$  and  $f_n^{-1}(w_{n+1})$  are within the balls around  $x_{n+1}, x_n$ , respectively, for which the lemmas 6.4–6.5 concerning growth bounds on  $f_n$  apply.

**Lemma 6.8.** The sequence  $\{w_n\}_{n=0}^N$  is an orbit of the dynamical system  $\{f_n\}$ . That is,  $f_n(w_n) = w_{n+1}$  for  $n = 0, \dots, N-1$ .

*Proof.* Equation (6) says that  $U_n(w_n - f_n^{-1}(w_{n+1})) = 0$ . Since  $S_n + U_n = I$ ,  $w_n - f_n^{-1}(w_{n+1})$  belongs to the subspace  $S_n$ . We evaluate  $|S_{n+1}(J_n(w_n - f_n^{-1}(w_{n+1})))|$  in two ways. First, using the fact that  $S_{n+1} + U_{n+1} = I$ ,

$$\begin{aligned} |S_{n+1}J_n(w_n - f_n^{-1}(w_{n+1}))| &\geq |J_n(w_n - f_n^{-1}(w_{n+1}))| - |U_{n+1}J_n(w_n - f_n^{-1}(w_{n+1}))| \\ &\geq \frac{1}{mB_1}|w_n - f_n^{-1}(w_{n+1})| - 2m^{3/2}B\delta|U_{n+1}||w_n - f_n^{-1}(w_{n+1})| \\ &= \left( \frac{1}{mB_1} - 2Bm^{3/2}|U_{n+1}|\delta \right) |w_n - f_n^{-1}(w_{n+1})| \end{aligned}$$

where the last inequality uses lemma 6.6.

On the other hand,

$$J_n(w_n - f_n^{-1}(w_{n+1})) = f_n(w_n) - w_{n+1} - (f_n(w_n) - w_{n+1} - Df_n(w_n)(w_n - f_n^{-1}(w_{n+1}))) \\ + (J_n - Df_n(w_n))(w_n - f_n^{-1}(w_{n+1})).$$

Since  $S_{n+1}(f_n(w_n) - w_{n+1}) = 0$  by equation (5), we have

$$|S_{n+1}(J_n(w_n - f_n^{-1}(w_{n+1})))| \\ \leq \frac{1}{2}m^{3/2}B_2|S_{n+1}||w_n - f_n^{-1}(w_{n+1})|^2 + m^{3/2}B_2|S_{n+1}||x_n - w_n||w_n - f_n^{-1}(w_{n+1})| \\ = m^{3/2}B_2|S_{n+1}|(\frac{1}{2}|w_n - f_n^{-1}(w_{n+1})| + 4\delta^{p+1/2})|w_n - f_n^{-1}(w_{n+1})|$$

Putting the two inequalities together, we have that either  $w_n = f_n^{-1}(w_{n+1})$ , in which case we are done, or else

$$\frac{1}{mB_1} - 2Bm^{3/2}|U_{n+1}|\delta \leq m^{3/2}B_2|S_{n+1}|(\frac{1}{2}|w_n - f_n^{-1}(w_{n+1})| + 4\delta^{p+1/2}) \\ \frac{1}{mB_1} \leq Bm^{3/2}\delta^p(2\delta^{1/2} + \frac{1}{2} + 4\delta^p)$$

where we use the bound  $\delta^{p-1/2}$  on  $|S_{n+1}|$  and  $|U_{n+1}|$ . This inequality implies that

$$\delta^{-p} \leq m^{5/2}B^2\left(\frac{1}{2} + \frac{2}{\sqrt{10m}} + \frac{4}{m^{5/2}B^2}\right) \\ < m^{5/2}B^2$$

since  $m \geq 2$ ,  $B \geq 2$ . This contradicts the assumption  $\delta^{-p} = m^{5/2}B^2$ . Therefore  $w_n = f_n^{-1}(w_{n+1})$  for  $n = 0, \dots, N-1$ .  $\square$

## References

- [1] Anosov D V 1967 Geodesic flows and closed Riemannian manifolds with negative curvature *Proc. Steklov Inst. Math.* **90**
- [2] Bowen R 1975  $\omega$ -limit sets for Axiom A diffeomorphisms *J. Diff. Eq.* **18** 333-9
- [3] Chow S-N and Palmer K 1990 On the numerical computation of orbits of dynamical systems: the one-dimensional case *Preprint*
- [4] Golub G and Van Loan C 1989 *Matrix Computations* 2nd edn (Baltimore, MD: The Johns Hopkins University Press)
- [5] Grebogi C, Hammel S and Yorke J 1987 Do numerical orbits of chaotic dynamical processes represent true orbits? *J. Complexity* **3** (1987) 136-45
- [6] Grebogi C, Hammel S and Yorke J 1988 Numerical orbits of chaotic processes represent true orbits *Bull. Am. Math. Soc.* **19** 465-70
- [7] Hammel S 1990 A noise reduction method for chaotic systems *Phys. Lett.* **148A** 421-8
- [8] Hartman P 1964 *Ordinary Differential Equations* (New York: Wiley)
- [9] Sauer T and Yorke J 1990 Shadowing trajectories in dynamical systems *Computer Aided Proofs in Analysis* ed K Meyer and D Schmidt (Berlin: Springer) pp 229-34

## Analysis of a procedure for finding numerical trajectories close to chaotic saddle hyperbolic sets<sup>†</sup>

HELENA E. NUSSE<sup>‡\*</sup> AND JAMES A. YORKE<sup>‡§</sup>

*University of Maryland, College Park, Maryland 20742, USA*

*(Received 1 February 1989 and revised October 1989)*

**Abstract.** In dynamical systems examples are common in which there are regions containing chaotic sets that are not attractors, e.g. systems with horseshoes have such regions. In such dynamical systems one will observe chaotic transients. An important problem is the 'Dynamical Restraint Problem': given a region that contains a chaotic set but contains no attractor, find a chaotic trajectory numerically that remains in the region for an arbitrarily long period of time.

We present two procedures ('PIM triple procedures') for finding trajectories which stay extremely close to such chaotic sets for arbitrarily long periods of time.

### 1. Introduction

Studying dynamical systems, one often observes transient chaotic behaviour, apparently due to the presence of horseshoes. For example, for suitably chosen parameter values, the Hénon map has an attracting period orbit with period 5 and also a non-attracting chaotic set, and one observes that the duration of the transient chaotic behaviour of many trajectories is rather short before they settle down on the period 5 attractor. Other famous examples with chaotic transients are: the Hénon map for large parameter values where almost all trajectories go to infinity and there is a bounded non-attracting invariant set; the forced damped pendulum; and the Lorenz equations for values of the Rayleigh number below the standard values that have a chaotic attractor. Transient chaos is also present whenever there is a fractal boundary separating the basins of two or more attractors.

Let  $M$  be a smooth  $n$ -dimensional manifold without boundary, and let  $F$  be a  $C^3$ -diffeomorphism from  $M$  to itself. We denote by  $\rho$  the distance function on  $M$ .

<sup>†</sup> Research in part supported by AFOSR, by DARPA under the Applied & Computational Mathematics Program, and the Netherlands Organization for the Advancement of Pure Research (N.W.O.).

<sup>‡</sup> Institute for Physical Science and Technology, University of Maryland.

<sup>\*</sup> Rijksuniversiteit Groningen, Fac. Economische Wetenschappen, WSN-gebouw, Postbus 800, NL-9700 AV Groningen, The Netherlands.

<sup>§</sup> Department of Mathematics, University of Maryland.

A region  $R$  is an open and bounded set in  $M$ . We say a region  $R$  is a *transient region* if it contains no attractor. We will be studying these regions in cases where the trajectory through almost every initial point eventually leaves the region. We investigate special trajectories that remain in such a transient region for all positive time. For example, the horseshoe is usually pictured mapping a rectangle to a horseshoe shape; the rectangle is a transient region. The great majority of the trajectories of the horseshoe map will leave the region after a few iterates. We are looking for numerical procedures for finding chaotic trajectories that stay in the transient region as long as we wish to compute them for  $t \geq 0$ . The main problem that we would like to address is:

*The dynamic restraint problem.* Find a (nonperiodic) orbit numerically that remains in a specified transient region for an arbitrarily long period of time.

The above problem explicitly concerns numerical (i.e. computer) procedures of finite precision. It leads to the following problem where it is assumed all computation can be made exactly.

*The static restraint problem.* Find an initial point whose orbit stays in a specified transient region for an arbitrarily long period of time.

We will establish a procedure (the *PIM triple procedure*) for finding points whose orbits will stay in specified regions in  $M$  for dynamical systems in ideal cases that are uniformly saddle-hyperbolic systems. The unstable manifold of each nonwandering point in the transient region is assumed to be one dimensional.

Let  $R$  be a transient region for  $F$ . The *stable set*  $S(R)$  of  $F$  is  $\{x \in R: F^n(x) \in R \text{ for } n = 0, 1, 2, \dots\}$ ; the *unstable set*  $U(R)$  of  $F$  is  $\{x \in R: F^{-n}(x) \in R \text{ for } n = 0, 1, 2, \dots\}$ . The set of points  $x$  for which  $F^n(x)$  is in  $R$  for all integers  $n$  is called the *invariant set*  $\text{Inv}(R)$  of  $F$  in  $R$ , that is,  $\text{Inv}(R) = S(R) \cap U(R)$ . A component of  $S(R)$  (resp.,  $U(R)$ ), which contains a point of  $\text{Inv}(R)$  is called a *stable* (resp., *unstable*) *segment*. We call  $\text{Inv}(R)$  a *chaotic saddle* when it includes a Cantor set. We assume that for the transient region  $R$  the set  $\text{Inv}(R)$  is nonempty.

We will refer to  $R \setminus S(R)$ , the complement of the stable set  $S(R)$  in the transient region  $R$ , as the *transient set*. We will say that a point  $p$  in  $S(R)$  is *accessible* from the transient set  $R \setminus S(R)$  if there is a continuous curve  $K$  ending at  $p$  so that  $K \setminus \{p\}$  is in the transient set  $R \setminus S(R)$ . For uses in dynamics, see [GOY] and [AY]. We would like to address the following problem:

*Accessible static restraint problem.* Given a segment  $J$  that intersects the stable set  $S(R)$  transversally, describe a procedure for finding a point (in  $J \cap S(R)$ ) which is accessible (from  $R \setminus S(R)$ ).

We will establish a procedure (the *Accessible PIM triple procedure*) for finding such accessible points in  $M$  for the same class of dynamical systems as above.

Both our procedures are based on our presumed ability to specify an initial point  $p$  and compute the time  $T_R(p)$  its trajectory takes to escape from  $R$ . In the PIM (Proper Interior Maximum) triple procedure, we seek out triples of points  $a$ ,  $c$ , and  $b$  on a curve segment with  $c$  the 'interior' point, that is,  $c$  is between  $a$  and  $b$ . The

triples are selected with an 'interior maximum' of the escape time, which means  $T_R(c) > T_R(a)$  and  $T_R(c) > T_R(b)$ . We then look for new triples that lie in the  $a, b$  segment but are closer together and so are 'proper'. The most challenging cases are those in which the average escape time is short so that the transient trajectories of typical points in  $R$  do not come close to the unstable chaotic set.

The organisation of the paper is as follows. In § 2 we present the PIM triple procedure and the Accessible PIM triple procedure; the main results for the validity of these procedures for hyperbolic systems are stated precisely in § 3. § 4 is devoted to the proofs of the results in § 3. In § 5, we will discuss the associated numerical procedures (including the shadowing of the numerical orbits by real orbits of the dynamical system). Finally in § 6, we will explain why the PIM triple methods also can be used for basin boundaries; we will describe how the results carry over to higher dimensional systems; and we also will argue that it is sufficient to assume that  $F$  is of class  $C^2$ .

## 2. The procedures

Let the manifold  $M$ , the diffeomorphism  $F$ , and the transient region  $R$  be as in the introduction.

The escape time  $T_R(x)$  of a point  $x$  in  $M$  for  $R$  is defined by

$$T_R(x) = \begin{cases} \min \{n \geq 0: F^n(x) \notin R\} \\ \infty & \text{if } F^n(x) \in R \text{ for all } n \geq 0. \end{cases}$$

For the example of the horseshoe map, the escape time function  $T_R$  has the following properties: (1)  $T_R(x) = \infty$  for  $x$  on a Cantor set of stable segments; (2) if  $a, c$ , and  $b$  are three points on a segment  $L$  of an unstable segment  $J$  so that: (i)  $c$  is between  $a$  and  $b$  and (ii)  $T_R(c) > \max \{T_R(a), T_R(b)\}$ , then the segment  $[a, b]_J \subset J$  from  $a$  to  $b$  intersects the stable set  $S(R)$ . These properties play a crucial role in the PIM triple procedures, and lead to the following definitions. Let  $J$  be an unstable segment in  $R$ . Then  $J$  is homeomorphic to an interval, and we may assume it has the ordering of an interval. The notation  $(a, c, b)$  for a triple means that  $a, c$ , and  $b$  lie on  $J$  and  $c$  is between  $a$  and  $b$ . Let  $L \subset J$  be a 'segment', that is, a connected subset of  $J$ . Assume  $L$  intersects the stable set  $S(R)$  transversally, and let  $(a, c, b)$  be a triple on  $L$ . Since  $L$  is homeomorphic to an interval it has an ordering. We assume that the ordering on  $J$  (and hence on  $L$ ) is such that  $a < c < b$ ; and for points  $x$  and  $y$  in  $J$  we write  $[x, y]_J$  for the segment on  $J$  joining  $x$  and  $y$ . The triple  $(a, c, b)$  is called an *Interior Maximum triple* if  $T_R(c) > \max \{T_R(a), T_R(b)\}$ ; and  $(a, c, b)$  is called a *Proper Interior Maximum (PIM) triple* on  $L$ , if  $(a, c, b)$  is an Interior Maximum triple and at least one of the points  $a$  and  $b$  is in the interior of  $L$ .

For each  $\varepsilon > 0$ , an  $\varepsilon$ -refinement of  $\{a, b\}$  (w.r.t.  $J$ ) is a finite set of points  $a = g_0 < g_1 < \dots < g_N = b$  in  $[a, b]_J$  such that

$$(\varepsilon/2) \cdot \rho([a, b]_J) \leq \rho([g_k, g_{k+1}]_J) \leq \varepsilon \cdot \rho([a, b]_J)$$

for all  $k$ ,  $0 \leq k \leq N-1$ , and an  $\varepsilon$ -refinement of  $(a, c, b)$  is an  $\varepsilon$ -refinement of  $\{a, b\}$  as above so that  $c = g_k$  for some  $k$ ,  $1 \leq k \leq N-1$ .

The outline of the PIM triple procedure is the following. Let  $R$  be an appropriately chosen transient region for  $F$ , and let  $L$  be a segment on an unstable segment  $J$  (intersecting the stable set transversally). Let  $\varepsilon > 0$  be sufficiently small. Given a PIM triple  $(a_n, c_n, b_n)$  in  $L$ , starting with  $n=0$ , choose some  $\varepsilon$ -refinement  $P_n$  of the triple  $(a_n, c_n, b_n)$  in  $[a_n, b_n]_J$ , select any three not necessarily consecutive points from  $P_n$  which constitute a new PIM triple  $(a_{n+1}, c_{n+1}, b_{n+1})$  on  $[a_n, b_n]_J$ . The new triple must be 'proper'; proper here means  $[a_{n+1}, b_{n+1}]_J$  is a proper subset of  $[a_n, b_n]_J$ . The condition guaranteeing the existence of such a PIM triple when  $\varepsilon$  is sufficiently small, will be described in § 3. Note that, according the definition of PIM triple,  $\rho([a_{n+1}, b_{n+1}]_J) \leq (1-0.5\varepsilon)\rho([a_n, b_n]_J)$ . Thus the nested sequence of the intervals  $\{[a_n, b_n]_J\}_{n \geq 0}$  converges to a point which we will call a *PIM limit point*. The  $\varepsilon$  above can be chosen small enough that it is independent of  $n$ . We will show that under reasonable conditions the orbit of the PIM limit point stays in the transient region  $R$ . The choice of the PIM triple is typically not unique and different choices will result in different PIM limit points. This 'static' problem's solution is not directly implementable on a computer because computations are made with finite precision, but it leads to a practical solution of the dynamic restraint problem as discussed in § 5.

The idea of the Accessible PIM triple procedure is like the PIM triple procedure except that the PIM triples  $(a_n, c_n, b_n)$  are selected more precisely so that  $[a_n, a_{n+1}]_J$  does not intersect the stable set  $S(R)$  for all  $n \geq N$  for some  $N \in \mathbb{N}$ . The difficulty here is that we only compute the escape times of the grid points and yet we must be sure that  $[a_n, a_{n+1}]_J$  contains no points of  $S(R)$ . We must guarantee the procedure will succeed if  $\varepsilon > 0$  is small enough, where  $\varepsilon$  is fixed, depending only on the diffeomorphism and region.

Our objective is to describe the *Accessible PIM triple procedure* that selects in a unique way a nested sequence of PIM triple intervals on  $J$  which leads to an accessible point in  $S(R)$  on  $J$ . The accessible point  $p$  in  $J \cap S(R)$  that we will find, will be accessible using the curve  $[r, p]_J$  for some  $r$  in  $J$ , so we say  $p$  will be 'accessed from the left', that is from the side containing  $r$ . We could alternatively have chosen to approach from the right and we would expect to find a different point.

Given an  $\varepsilon/3$ -refinement  $P_n = \{x_i : 0 \leq i \leq N(\varepsilon)\}$  on  $J$  of a PIM triple  $(a_n, c_n, b_n)$  in  $J$  with  $a_n = x_0 < x_1 < \dots < x_{N(\varepsilon)} = b_n$ . Assuming that  $P_n$  includes a PIM triple, then we choose the next PIM triple  $(a_{n+1}, c_{n+1}, b_{n+1})$  in  $P_n$  in the following way:

(1) Select  $b_{n+1}$  to be the leftmost point in  $P_n$  such that it is the right point of a PIM triple in  $P_n$ :

(2) Select  $c_{n+1}$  to be the adjacent point to the left of  $b_{n+1}$  in  $P_n$ ;

(3) The systematic choice of  $a_{n+1}$  in  $P_n$  is the following: Let  $M_n$  be the minimum value of  $\{T_R(x_i) : x_i \in P_n, x_i < c_{n+1}\}$ . We write:

$a_{n+1}^0$  is the rightmost point of  $\{x_i \in P_n : x_i < c_{n+1}, T_R(x_i) = M_n\}$ ;

$a_{n+1}^+$  is the adjacent point to the right of  $a_{n+1}^0$  in  $P_n$ ;

$a_{n+1}^1$  is the rightmost point of  $\{x \in P_n : x \leq c_{n+1}, T_R(x) = T_R(a_{n+1}^-)\}$ .

Case (i). If either  $M_n < T_R(a_n)$  or  $M_n > \min \{T_R(x) : x \in P_n\}$ , then choose  $a_{n+1} = a_{n+1}^0$ ; otherwise,

Case (ii). If  $M_n = T_R(a_n)$  and  $P_n$  is not an  $\varepsilon$ -refinement of  $(a_n, c_{n+1}, b_{n+1})$ , then choose  $a_{n+1} = a_n$ ; otherwise,

Case (iii). If  $M_n = T_R(a_n)$  and  $P_n$  is an  $\varepsilon$ -refinement of  $(a_n, c_{n+1}, b_{n+1})$ , and if  $a_{n+1}^0 > a_n$  or  $a_{n+1}^1 = c_{n+1}$ , then choose  $a_{n+1} = a_{n+1}^0$ ; otherwise,

Case (iv). If  $M_n = T_R(a_n)$  and  $P_n$  is an  $\varepsilon$ -refinement of  $(a_n, c_{n+1}, b_{n+1})$ , and if  $a_{n+1}^0 = a_n$  and  $a_{n+1}^1 < c_{n+1}$ , then choose  $a_{n+1} = a_{n+1}^1$ .

Repeatedly applying the Accessible PIM triple procedure leads to an accessible point on  $S(R)$ .

To understand rule 3, notice that rules 1 and 2 imply that the graph of  $T_R$  is rather simple on  $P_n \cap [a_n, c_{n+1}]$ , namely,  $T_R$  is monotonic increasing on  $P_n$  between  $a_{n+1}^0$  and  $c_{n+1}$ , and  $T_R$  is non-increasing on  $P_n$  between  $a_n$  and  $a_{n+1}^0$ . These properties follow from the fact that  $b_{n+1}$  was chosen as far left as possible. We will show that after the first few iterates  $T_R(a_n) = \min \{T_R(x) : x \in P_n\}$ .

### 3. Results

In § 2 we presented the idea of the procedure for finding a point whose orbit stays in the transient region. In this description, we assumed that there exists an  $\varepsilon > 0$  for which every  $\varepsilon$ -refinement of a PIM triple includes a new PIM triple. Furthermore, the associated curve segment from  $a_{n+1}$  to  $b_{n+1}$  has a length at most  $(1 - \varepsilon/2)$  times the length of the previous one (from  $a_n$  to  $b_n$ ). We will justify these concepts.

Let the manifold  $M$  and diffeomorphism  $F$  be as in the introduction. A subset  $\Lambda$  of  $M$  is *hyperbolic* if it is closed and  $F$ -invariant and the tangent bundle  $T_\Lambda M$  splits into  $dF$ -invariant subbundles  $E^s$  and  $E^u$  on which  $dF$  is uniformly contracting and uniformly expanding respectively. A hyperbolic set  $\Lambda$  is called a *saddle-hyperbolic set* if  $\dim E^s \geq 1$  and  $\dim E^u \geq 1$ . We will call a region  $R$  a *saddle-hyperbolic transient region* if  $R$  satisfies all the following conditions:

- (A1)  $R$  is a transient region;
- (A2) *Hyperbolicity property*:  $\text{Inv}(R)$  is a nonempty saddle-hyperbolic set;
- (A3) *Boundary property*:  $\overline{U(R)} \cap \partial R$  is mapped outside the closure of  $R$ ;
- (A4) *Intersection property*: each nontrivial component  $\gamma$  of  $U(R)$  is an unstable segment, that is,  $\gamma$  intersects  $\text{Inv}(R)$ ; note that such a segment  $\gamma$  must intersect  $S(R)$  transversally.

We assume throughout that  $\dim E^u = 1$ . For the sake of simplicity, we assume that  $n = 2$ ; the more difficult case  $n \geq 3$  will be discussed in § 6.

For a saddle-hyperbolic transient region  $R$  and  $\varepsilon > 0$ , the properties (A1) and (A2) imply that the escape time of almost every point on an unstable segment is finite. (A result due to Bowen and Ruelle [BR] shows that  $S(R)$  has Lebesgue measure zero.) Hence, one may assume that such a refinement does not intersect the stable set  $S(R)$ .

If  $R$  is a saddle-hyperbolic transient region, then the escape time map  $T_R$  restricted to an unstable segment  $J \subset U(R)$  has the following two properties, which follow from Proposition 1 and the  $T$ -Jump Lemma below.

- (i) All the points in a chosen segment  $[a, b]$ , on  $J$  will escape from  $R$  if and only if no  $\varepsilon$ -refinement of  $\{a, b\}$  includes a PIM triple;
- (ii)  $T_R$  is locally constant on an open subset of full measure of  $J$ , and if  $T_R(x) < \infty$  and  $x$  is a point of discontinuity of  $T_R$ , then

$$\liminf_{y \rightarrow x} T_R(y) = T_R(x) \quad \text{and} \quad \limsup_{y \rightarrow x} T_R(y) = T_R(x) + 1.$$

*Application.* The objective of the paper is to present procedures which enable us to obtain numerical trajectories lying on chaotic saddles, and to justify that these procedures work in ideal cases. The examples of interest will rarely satisfy all our hypotheses, and yet we observe that frequently we can successfully use the procedures to obtain pictures of  $\text{Inv}(R)$  by plotting the numerical trajectory. Consider the following example.

Let the diffeomorphism  $F$  acting on the plane be given by

$$F(x, y) = (A - x^2 + M \cdot y, x).$$

It is well known that the map  $F$  is equivalent under a linear change of variables with the Hénon map. We choose the parameter values  $M = 0.3$ , and  $A = 3$  in figure 1(a),  $A = 4.2$  in figure 1(b) (and figure 2) and  $A = 2.0$  in figure 1(c). Then a result due to Devaney and Nitecki [DN] implies that  $B = \{(x, y) : -3 < x < 3, -3 < y < 3\}$  includes the nonwandering set of  $F$ , so we select  $B$  for the transient region. When  $A = 4.2$ , the nonwandering set is a uniformly hyperbolic chaotic saddle. We start the numerical procedure with the horizontal line segment with  $y = 1$  extending from the left side of  $B$  to the right side. By using the 'PIM triple procedure' we obtain a numerical trajectory consisting of tiny intervals. The result is presented in figure 1.

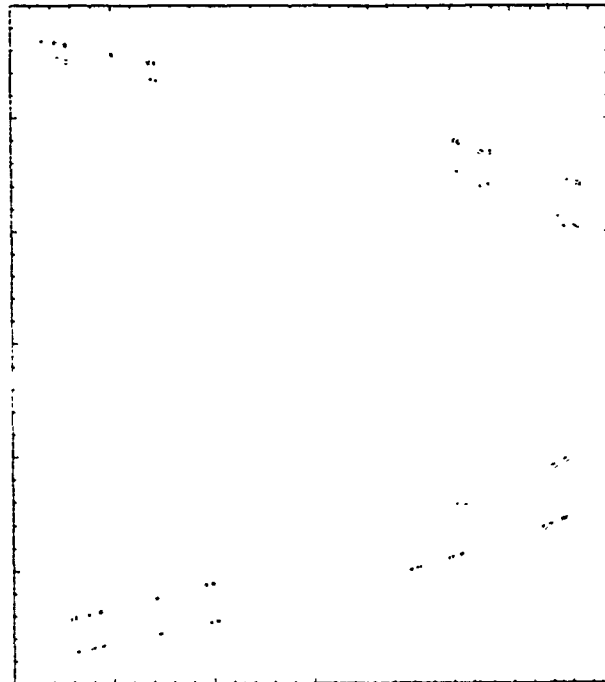
When  $A = 4.2$ , the region  $B$  is a saddle-hyperbolic transient region: the results due to Devaney and Nitecki [DN] imply that  $B$  satisfies the conditions (A1)–(A4). When  $A = 3$  we do not know if condition (A2) will hold, and for  $A = 2.0$  we have a non-fully developed horseshoe.

In figure 2 we present the sets  $U(B)$  and  $S(B)$  for  $A = 4.2$  (the chaotic saddle is the intersection  $S(B) \cap U(B)$ ), and the accessible fixed point on the chaotic saddle is indicated by an arrow.

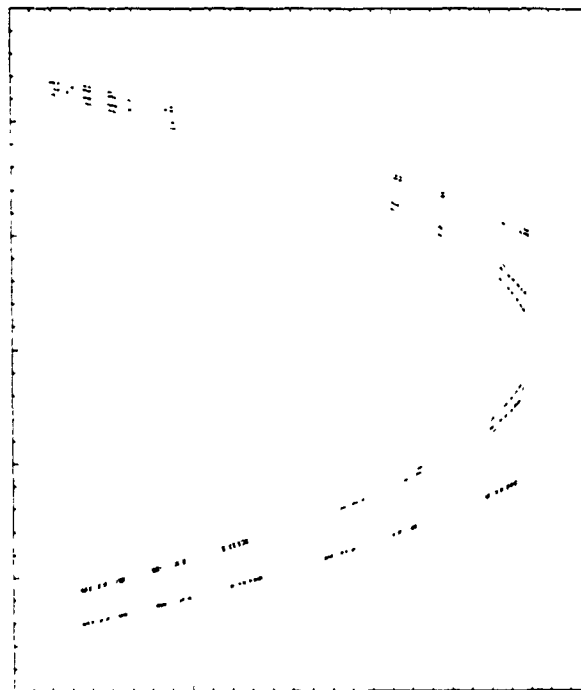
The reader is referred to [NY] for other applications such as the Lorenz equations, the pulsed rotor map, and the forced pendulum equation.

Rather than state one or two theorems the results seem best stated a progression of ideas: (1) PIM triples exist, (2) refinement of PIM triples include PIM triples, and (3) the resulting sequence of PIM triples converge to a desirable point. The special case of accessible PIM triple sequences must be discussed separately.

From now on, we will assume that  $R$  is a saddle-hyperbolic transient region for  $F$  with  $\dim E^u = 1$ , and that  $J \subset U(R)$  denotes an unstable segment. That implies that both ends of  $J$  are in the boundary of the transient region  $R$ . We know by the Intersection assumption that  $J$  intersects the stable set  $S(R)$ . Clearly, this property



(a)



(b)

FIGURE 1. (a) Numerical trajectory obtained by the PIM triple procedure for the Hénon map in the transient region  $-3 < x, y < 3$ , and parameter values  $A=3$ ,  $M=0.3$ . (b) Numerical trajectory obtained by the PIM triple procedure for the Hénon map in the transient region  $-3 < x, y < 3$ , and parameter values  $A=4.2$ ,  $M=0.3$ . (c) Numerical trajectory obtained by the PIM triple procedure for the Hénon map in the transient region  $-3 < x, y < 3$ , and parameter values  $A=2.0$ ,  $M=0.3$ .

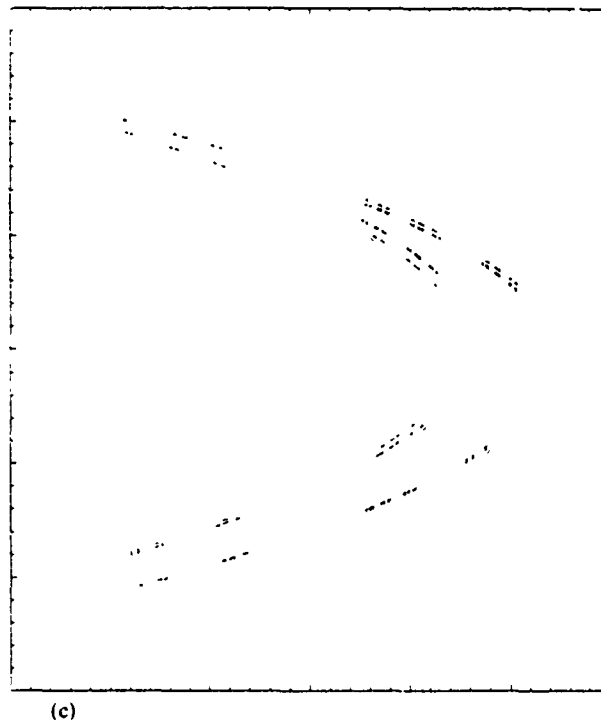


FIGURE 1—continued.

will not hold for each subsegment  $L$  of  $J$ , since  $J \cap S(R)$  is nowhere dense in  $J$  and one can choose the segment  $L$  lying entirely in the complement of  $J \cap S(R)$ . Our first result 'PIM Existence Proposition' characterizes the segments intersecting the stable set  $S(R)$ .

**PROPOSITION 1. (PIM existence.)** *Let  $L = [a, b]$ , be a segment in  $J$ . The following statements are equivalent:*

- (i) *there exists  $\varepsilon > 0$  such that every  $\varepsilon$ -refinement of  $\{a, b\}$  includes a PIM triple;*
- (ii)  *$L$  contains a point of  $\text{Inv}(R)$  in its interior.*

In the PIM Existence Proposition the segment  $L$  can be chosen so that it intersects  $S(R)$  only at points extremely close to one of the end points of  $L$  and so  $\varepsilon$  must be extremely small, so  $\varepsilon$  depends on the choice of  $L$ . However, the PIM Refinement Proposition, stated below, shows that a single  $\varepsilon$  (depending on  $F$  and  $R$ ) can be used, once we have found our first PIM triple. One might expect that our assumptions of uniform hyperbolicity would imply that the uniformity of  $\varepsilon$  would be an easy corollary. In fact, the existence of an  $\varepsilon$  for each PIM triple is much easier than  $\varepsilon$  can be chosen uniformly, and this uniformity is essential for the PIM triple procedures. In principle it can be difficult to find the first PIM triple if the initial interval  $L$  is chosen badly.

**PROPOSITION 2. (PIM refinement.)** *There exists  $\varepsilon > 0$  (depending on  $F$  and  $R$ ) such*

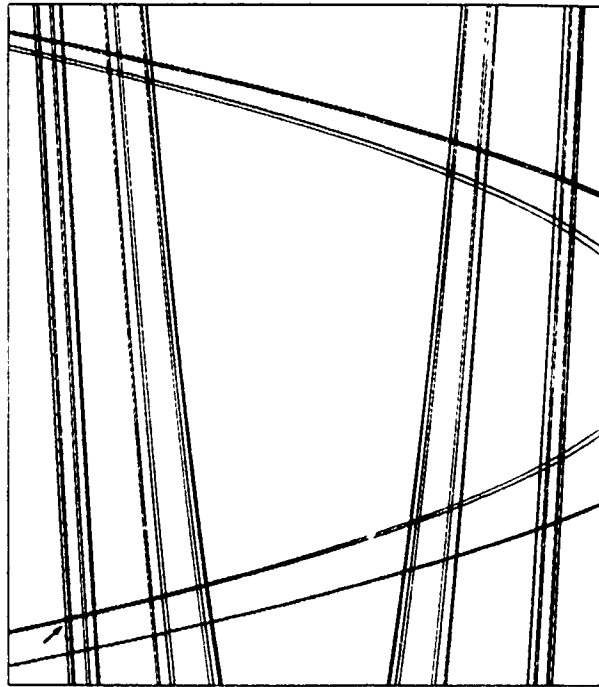


FIGURE 2. The stable and the unstable manifold for the fixed point at approx. (1.729, 1.729) for the Hénon map in the transient region  $-3 < x, y < 3$ , and parameter values  $A = 4.2$ ,  $M = 0.3$ . The accessible fixed point on the chaotic saddle is indicated by an arrow.

that there is a PIM triple in each  $\varepsilon$ -refinement in  $J$  of each Interior Maximum triple in  $J$ , for every unstable segment  $J \subset U(R)$ .

The next result deals with the convergence of the sequence of nested PIM triple segments  $[a_{n+1}, b_{n+1}]_J \subset [a_n, b_n]_J$  on  $J$ , in other words, the PIM triple procedure is valid. A sequence of PIM triples  $\{(a_n, c_n, b_n)\}_{n \geq 0}$  on  $J$  is called a *PIM triple sequence* if  $(a_{n+1}, c_{n+1}, b_{n+1})$  is in an  $\varepsilon$ -refinement of the Interior Maximum triple  $(a_n, c_n, b_n)$  for all  $n$ . We say  $\{(a_n, c_n, b_n)\}_{n \geq 0}$  is the *accessible PIM triple sequence* if  $(a_n, c_n, b_n)$  is selected using the Accessible PIM triple procedure for all  $n$ .

**PROPOSITION 3. (PIM convergence.)** *Let  $\varepsilon > 0$  be as in Proposition 2. Every sequence of nested segments  $\{[a_n, b_n]_J\}_{n \geq 0}$  that is associated with the PIM triple sequence  $\{(a_n, c_n, b_n)\}_{n \geq 0}$  on  $J$ , converges to a point on  $S(R)$ .*

The next result is the key in proving that the 'Accessible PIM triple procedure' is valid.

**PROPOSITION 4 (Accessible PIM Refinement.)** *Let  $\varepsilon > 0$  be as in Proposition 2. Let  $\{(a_n, c_n, b_n)\}_{n \geq 0}$  be an Accessible PIM triple sequence on  $J$ . Then there exists integer  $N \geq 0$  such that  $[a_n, a_{n+1}]_J$  does not intersect  $S(R)$  for every  $n \geq N$ .*

Recall that a nested sequence of PIM triple intervals obtained from  $\varepsilon$ -refinements will converge to a PIM limit point on  $S(R)$ . Note that the PIM limit point of the PIM triple intervals associated with PIM triples in Proposition 4 is an accessible

point on  $S(R)$ . The next result implies that the Accessible PIM triple procedure is valid.

**PROPOSITION 5.** (Accessible PIM convergence.) *Let  $\varepsilon > 0$  be as in Proposition 2. If the PIM triple sequence  $\{(a_n, c_n, b_n)\}_{n \geq 0}$  in Proposition 3 is an accessible PIM triple sequence, then the sequence of nested segments  $\{[a_n, b_n]\}_{n \geq 0}$  on  $J$ , converges to an accessible point on  $S(R)$ .*

#### 4. Proofs

##### 4.1. Preliminaries

We assume that  $R$  is a saddle-hyperbolic region for the diffeomorphism  $F$ . By Smale's 'Spectral Decomposition Theorem' [S] we know that we can decompose the nonwandering set  $\Omega$  into a finite collection of disjoint closed invariant subsets and on each of these subsets  $F$  has a dense orbit; these maximal invariant subsets of  $\Omega$  appearing in the decomposition are called the *basic sets* (see e.g. [Ni] and [GH] for the definitions and several properties of uniformly hyperbolic systems). From now on, let  $\Gamma$  denote a basic set of  $F$ . From the definition of  $\text{Inv}(R)$  it follows immediately that either  $\Gamma \subset \text{Inv}(R)$  or  $\Gamma \cap \text{Inv}(R) = \emptyset$ . This implies that we can decompose  $\text{Inv}(R)$  into finitely many basic sets. Note that ' $\Gamma \cap \text{Inv}(R) = \emptyset$ ' does not imply ' $\Gamma \cap R = \emptyset$ ', and ' $\Gamma \cap R \neq \emptyset$ ' does not imply ' $\Gamma \cap \text{Inv}(R) \neq \emptyset$ '.

Recall that for  $z \in \Omega$  the stable manifold  $W^s(z)$  (resp. unstable manifold  $W^u(z)$ ) of  $z$  is the set of points  $x$  for which  $\rho(F^n(z), F^n(x)) \rightarrow 0$  (resp.  $\rho(F^{-n}(z), F^{-n}(x)) \rightarrow 0$ ) as  $n \rightarrow \infty$ ; the local stable manifold  $W_{\text{loc}}^s(z)$  (resp. the local unstable manifold  $W_{\text{loc}}^u(z)$ ) of  $z$  of size  $\beta$  is the set of points  $x$  in  $W^s(z)$  (resp.  $W^u(z)$ ) so that  $\rho(F^n(z), F^n(x)) \leq \beta$  (resp.  $\rho(F^{-n}(z), F^{-n}(x)) \leq \beta$ ) for all integers  $n \geq 0$ , where  $\beta > 0$ . When the stable or unstable manifold is a curve, we write  $W_{\text{loc}}^{\sigma+}(z)$  and  $W_{\text{loc}}^{\sigma-}(z)$  for the two components of  $W_{\text{loc}}^{\sigma}(z) \setminus \{z\}$ , where  $\sigma$  is either  $s$  or  $u$ .

We will call  $\Gamma$  a *trivial* basic set if  $\Gamma$  consists of one periodic orbit, and we call  $\Gamma$  a *nontrivial* basic set if  $\Gamma$  includes more than one periodic orbit. Assume that  $\Gamma$  is nontrivial, we call  $\Gamma$  *periodic* if there exists  $m \in \mathbb{N}$  such that  $F^m$  has no dense orbit on  $\Gamma$ , and we call  $\Gamma$  *nonperiodic* if it is not periodic. The following results 4.1, 4.2, and 4.4 are reformulated from [NP] and [PT].

**PROPOSITION 4.1.** *There exists finite sets  $P$ ,  $P^s$ , and  $P^u$  of periodic points,  $P = P^s \cup P^u$ , such that for all  $x \in \text{Inv}(R)$ :*

- (1) *If  $x$  is not a limit point of both  $W_{\text{loc}}^{u+}(x) \cap \Omega$  and  $W_{\text{loc}}^{u-}(x) \cap \Omega$ , then  $x$  is in  $W^s(p)$  for some  $p \in P^u$ .*
- (2) *If  $x$  is not a limit point of both  $W_{\text{loc}}^{s+}(x) \cap \Omega$  and  $W_{\text{loc}}^{s-}(x) \cap \Omega$ , then  $x \in W^u(p)$  for some  $p \in P^s$ .*

*Proof.* For a proof, see Newhouse and Palis [NP].

**PROPOSITION 4.2.** *Let  $P^s$  and  $P^u$  be as in Proposition 4.1. Let  $\Gamma$  be a nontrivial nonperiodic basic set in  $\text{Inv}(R)$ . Then there exist finitely many disjoint regions  $R_i$ , being diffeomorphic images of the square  $B = [-1, 1] \times [-1, 1]$ , say  $R_i = g_i(B)$ ,  $1 \leq i \leq N$  for some  $N \in \mathbb{N}$ , such that: (1)  $\Gamma \cap R_i \neq \emptyset$  for all  $i$ ; (2)  $\Gamma \subset \bigcup_{i=1}^N R_i$ ; (3)  $F(\partial_i R_i) \subset \bigcup_{i=1}^N \partial_i R_i$ , and  $F^{-1}(\partial_u R_i) \subset \bigcup_{i=1}^N \partial_u R_i$ , where  $\partial_i R_i = g_i(\{(x, y) : |x| = 1, -1 \leq y \leq 1\})$  resp.*

$\partial_u R_i = g_i(\{(x, y): -1 \leq x \leq 1, |y| = 1\})$  are segments in the stable set  $W^s(P^u)$  resp. the unstable set  $W^u(P^s)$ .

*Proof.* For a proof, see Palis and Takens [PT].

*Remark.* The intersection of  $\Gamma$  with the union of the regions in Proposition 4.2 is a Markov partition for  $\Gamma$ , see Bowen [B] for the notion of Markov partition.

**PROPOSITION 4.3.** *Let  $P^u$  be as in Proposition 4.1. Then we have  $x \in S(R)$  is accessible if and only if  $x \in W^s(p)$  for some  $p \in P^u$ .*

*Proof.* Apply the Propositions 4.1 and 4.2.

From now on, let  $z \in \Gamma \subset \text{Inv}(R)$  be fixed, and let  $I^u \subset W^u(z)$  be a segment such that  $I^u$  crosses each region  $R_k$  at least once, where  $R_k$ ,  $1 \leq k \leq N$ , is as in Proposition 4.2. Palis and Takens [PT] have shown that there exist finitely many disjoint regions denoted  $R_j(I^u)$  in  $\bigcup_{i=1}^N R_i$  that have the same properties as the  $R_i$ 's such that  $I^u$  crosses each  $R_j(I^u)$  exactly once,  $1 \leq j \leq N(I^u)$ , for some  $N(I^u) \in \mathbb{N}$ . Therefore, we will assume that  $I^u$  crosses each region  $R_i$  from Proposition 4.2 precisely once. There exist a  $C^{1+\alpha}$  stable foliation  $\tilde{\gamma}^s$  on a neighborhood  $U_\Gamma$  of  $\Gamma$  for some  $\alpha > 0$ , and it is no restriction to assume that every region  $R_i$  is contained in  $U_\Gamma$ ,  $1 \leq i \leq N$ ; see [PT].

Let  $\tau: \mathbb{R} \rightarrow W^u(z)$  be a  $C^3$  parametrization, and define a projection  $\pi: \Gamma \rightarrow \bigcup_{i=1}^N R_i \cap I^u$  by taking in each region  $R_i$ ,  $1 \leq i \leq N$ , the projection along the local stable manifolds into the intersection  $I^u$  with that region. This projection can be extended from  $\Gamma$  to the union of the regions  $R_i$ , by projecting along the leaves of the foliation  $\tilde{\gamma}^s$ . This extension will also be denoted by  $\pi$ . We obtain (see [PT]) the following result that says that for some iterate  $M$ , the map  $F$  can be viewed as expansive along unstable segments.

**PROPOSITION 4.4.** *There exist a positive integer  $M$  and a  $C^{1+\alpha}$  map  $\varphi: \bigcup_{i=1}^N \tau^{-1}(I^u \cap R_i) \rightarrow \mathbb{R}$  defined by  $\varphi(x) = \tau^{-1} \circ \pi \circ F^M \circ \tau(x)$  such that  $|\varphi'(x)| > 1$ , for some  $\alpha > 0$ .*

From now on, let  $I_1, \dots, I_N$  be  $N$  disjoint compact intervals on the real line, and we write  $Y = \bigcup_{j=1}^N I_j$ . Let  $f: Y \rightarrow \mathbb{R}$  be a  $C^{1+\alpha}$  map, for some  $\alpha > 0$ , with the following properties:

- (1)  $f$  is  $C^{1+\alpha}$  on an open neighborhood  $U$  of  $Y$ ;
- (2)  $Y \subset \text{Interior}(f(Y))$ ;
- (3) there exists numbers  $\lambda_j > 1$  such that  $|f'(x)| \geq \lambda_j$  for every  $x \in I_j$ ,  $1 \leq j \leq N$ ;
- (4) the transition matrix  $A_{Y,f}$  is primitive, that is, there is an integer  $Q > 0$  so that all the entries of  $(A_{Y,f})^Q$  are positive; where  $A_{Y,f}$  is defined by  $A_{Y,f}(j, m) = 1$  if  $f(I_j) \cap I_m \neq \emptyset$ , and  $A_{Y,f}(j, m) = 0$  if  $f(I_j) \cap I_m = \emptyset$ ,  $1 \leq j, m \leq N$ .

Note that condition (2) implies that either  $I_j \cap f(I_m) = \emptyset$  or  $I_j \cap \text{Int}(f(I_m)) = I_j$ , for all  $1 \leq j, m \leq N$ .

The escape time  $T_Y(x)$  of  $x \in Y$  under  $f$  is the minimum value  $n$  with the property  $f^n(x)$  is not in  $Y$ . We define for every integer  $k \geq 1$ :

$$A_k = \{x \in Y: T_Y(x) \geq k\}$$

$$D_k = \{x \in Y: T_Y(x) = k\}.$$

In particular,  $A_1 = Y$ . Hence, for each integer  $k \geq 1$  we have  $A_{k+1}$  is the set of points in  $A_k$  whose escape time from  $Y$  is at least  $k$ ; hence  $A_{k+1}$  is the set of points in  $Y$  that stay in  $Y$  under  $f^k$ . The points in  $Y$  which will stay in  $Y$  under all iterates will be denoted by

$$A_\infty = \{x \in Y : T_Y(x) = \infty\}$$

For every integer  $k \geq 1$  we have:

- (a)  $A_k = A_{k+1} \cup D_k$ ;
- (b)  $Y = A_{k+1} \cup \bigcup_{j=1}^k D_j$ , that is,  $Y$  is the union of the set of points  $A_{k+1}$  whose escape time from  $Y$  is at least  $k+1$ , and the set of points  $D_j$  whose escape time from  $Y$  is  $j$ , where  $1 \leq j \leq k$ .

Denote the length of an interval  $L$  by  $|L|$ .

**GEOMETRIC LEMMA I.** *There exists  $\delta_f > 0$  such that for every integer  $k \geq 1$ , the following holds:*

- (i) *Every component of  $A_k$  contains components of  $D_k$  and of  $A_{k+1}$ ;*
- (ii) *For each component  $D$  of  $D_k \cap A$ , one has  $|D|/|A| \geq \delta_f$ , and each component  $U$  of  $A_{k+1} \cap A$ , satisfies  $|U|/|A| \geq \delta_f$ , with  $A$  an arbitrarily chosen component of  $A_k$ .*

*Proof of the Geometric Lemma I.* For each integer  $i \geq 1$ , we write  $R_i$  for the sum of the entries on the  $i$ th row of  $A_{Y,j}$ ,  $1 \leq i \leq N$ . Assumption (4) implies  $R_i$  is at least 1 for all  $i$ , and the sum of the  $R_i$ 's is greater than  $N$ .

*Proof of (i).* Let  $k \geq 1$  be a given integer. First, we assume  $k = 1$ . Let  $L$  be a given component of  $A_1 = Y$ , say  $L = I_j$  for some  $j$ ,  $1 \leq j \leq N$ . The assumptions (1)-(4) imply  $f(L)$  contains  $R_j + 1$  components of  $D_1$ . Since  $L = \{x \in L : T_Y(x) \geq 2\} \cup \{x \in L : T_Y(x) = 1\}$ , we have that  $L$  contains  $R_j$  components of  $A_2$ .

Now we assume  $k > 1$ . Let  $A$  be a given component of  $A_k$ . By the definition of  $A_k$  and the assumptions on  $f$ , we have  $f^{k-1}(A)$  is a component of  $A_1$ , say  $f^{k-1}(A) = I_j$  for some  $j$ ,  $1 \leq j \leq N$ . Therefore,  $A$  contains  $R_j + 1$  components of  $D_k$  and  $R_j$  components of  $A_{k+1}$ .

*Proof of (ii).* We are looking for  $\delta_f > 0$  such that for each integer  $k \geq 1$  and for each component  $A$  in  $A_k$ , we have  $\mu(D)/\mu(A) \geq \delta_f$ , and  $\mu(U)/\mu(A) \geq \delta_f$ , for each component  $D$  of  $D_k \cap A$ , and each component  $U$  of  $A_{k+1} \cap A$ .

From (i) and the assumptions on  $f$  we obtain that for each  $k \geq 1$ , the number of components of  $A_k$  and that of  $D_k$  is finite. Let, for each integer  $k \geq 1$ ,  $N(A_k)$  be the number of components of  $A_k$ , and let  $N(D_k)$  be the number of components of  $D_k$ . We write, for each  $k \geq 1$ , the sets  $A_k$  and  $D_k$  as the union of their components in the following way:

$$A_k = \bigcup_{i=1}^{N(A_k)} A_{k,i}, \quad D_k = \bigcup_{i=1}^{N(D_k)} D_{k,i}.$$

For each  $k \geq 1$  and each component  $A$  in  $A_k$ , we define

$$\delta_k(A) = \min_V |V|/|A|,$$

where the minimum is taken over all components  $V$  of the sets  $D_k$  and  $A_{k+1}$ ; and we define

$$\delta_k = \min_A \delta_k(A),$$

where the minimum is taken over all components  $A$  of the set  $A_k$ . We are done if there exists  $\delta_f > 0$  so that  $\delta_k \geq \delta_f$  for all  $k$ .

Let  $k > 1$  be a given integer. Let  $A$  be a given component of  $A_k$ , and let  $D$  be an arbitrary component of either  $D_k$  or  $A_{k+1}$  such that  $A$  includes  $D$ . From the foregoing, we can fix an integer  $n(k)$ ,  $1 \leq n(k) \leq N(A_k)$ , such that  $A = A_{k,n(k)}$ , and an integer  $m(k)$ ,  $1 \leq m(k) \leq N(D_k)$  if  $D$  is a component of  $D_k$ , and  $1 \leq m(k) \leq N(A_{k+1})$  if  $D$  is a component of  $A_{k+1}$ , such that  $D = D_{k,m(k)}$ .

Set for each integer  $i$ ,  $2 \leq i \leq k$ :

$$A_{i-1,n(i-1)} = f(A_{i,n(i)}) \quad D_{i-1,m(i-1)} = f(D_{i,m(i)}).$$

Applying the mean value theorem, we can find for every integer  $i$ ,  $2 \leq i \leq k$ , real numbers  $a_i$  in  $A_{i,n(i)}$  and  $d_i$  in  $D_{i,m(i)}$  such that  $|f'(a_i)| \cdot |A_{i,n(i)}| = |A_{i-1,n(i-1)}|$ ,  $|f'(d_i)| \cdot |D_{i,m(i)}| = |D_{i-1,m(i-1)}|$ . This leads to:

$$|D_{k,m(k)}|/|A_{k,n(k)}| = \left( \prod_{i=2}^k |f'(a_i)/f'(d_i)| \right) \cdot (|D_{1,m(1)}|/|A_{1,n(1)}|). \quad (1)$$

From now on, we can mimic the proof of Lemma 5.5 in [Nu], and we obtain:

$$\lim_{k \rightarrow \infty} \prod_{i=2}^k |f'(a_i)/f'(d_i)| > 0. \quad (2)$$

The results (1) and (2) imply that there exists  $\gamma > 0$  such that  $|D_{k,m(k)}|/|A_{k,n(k)}| \geq \gamma$ . Therefore,  $\delta_k \geq \gamma$  for each  $k \geq 2$ .

We conclude:  $|D|/|A| \geq \delta_f$  for every component  $A$  in  $A_k$ , for every component  $D$  with  $D \subset A$ , where  $D$  is either a component of  $D_k$  or  $D$  is a component of  $A_{k+1}$ , for  $\delta_f = \min \{\delta_1, \gamma\}$ . This completes the proof of the Geometric Lemma 1.

#### 4.2. Proofs of the PIM propositions

Let  $J \subset U(R)$  denote an unstable segment. Recall that both end points are on the boundary of the transient region  $R$ , and that  $J$  intersects the stable set  $S(R)$ .

We define for every integer  $k \geq 1$ :

$$A_k(J) = \{x \in J : T_R(x) \geq k\} \quad D_k(J) = \{x \in J : T_R(x) = k\}.$$

In particular,  $A_1(J) = J$ . Hence, for each integer  $k \geq 1$  we have  $A_{k+1}(J)$  is the set of points in  $A_k(J)$  whose escape time from  $R$  is at least  $k+1$ ; hence,  $A_{k+1}(J)$  is the set of points in  $J$  that stay in  $R$  under  $F^k$ . The points in  $J$  which will stay in  $R$  under all iterates will be denoted by  $A_\infty(J)$ . For every integer  $k \geq 1$  we have:

$$(a) \quad A_k(J) = A_{k+1}(J) \cup D_k(J) \quad (b) \quad J = A_{k+1}(J) \cup \bigcup_{j=1}^k D_j(J),$$

that is,  $J$  is the union of the set of points  $A_{k+1}(J)$  the escape time of which from  $R$  is at least  $k+1$ , and the set of points  $D_j(J)$  the escape time of which from  $R$  is  $j$ , where  $1 \leq j \leq k$ . We write  $D_\infty(J) = \bigcup_{k=1}^\infty D_k(J)$ . Note that  $A_\infty(J) = \bigcap_{k=0}^\infty A_k(J)$ , and  $J = A_\infty(J) \cup D_\infty(J)$ . In the lemma below we will state that, if the value of the escape time map  $T_R$  changes then it changes in steps of 1.

**T-JUMP LEMMA.** For every  $x \in D_\infty(J)$  there exists an  $\varepsilon > 0$  such that for each  $y \in J$  with  $\rho([x, y]_J) < \varepsilon$  one has  $|T_R(x) - T_R(y)| \leq 1$ .

**Proof of the T-Jump Lemma.** Let  $x \in D_\infty(J)$  be given. We will write  $D_\infty^{\text{int}}(J) = \bigcup_{k=1}^\infty \text{Int}(D_k(J))$ , where  $\text{Int}(D_k(J))$  means the interior of  $D_k(J)$  for each  $k \geq 1$ . First, consider the case where  $x \in D_\infty^{\text{int}}(J)$ . Then, by the definitions,  $T_R$  is constant on the component of  $D_\infty^{\text{int}}(J)$  including  $x$ . Consequently, there exists an  $\varepsilon > 0$  so that  $T_R(y) = T_R(x)$  for all  $y$  in  $J$  with  $\rho([x, y]_J) < \varepsilon$ .

Now we consider the case where  $x \in D_\infty(J) \setminus D_\infty^{\text{int}}(J)$ . Let  $M \geq 0$  be the integer for which  $F^M(x) \in \text{Bndy}(R)$ , where  $\text{Bndy}(R)$  means the boundary of  $R$ . From the fact that each point in  $\text{Bndy}(R)$  is mapped outside  $\bar{R}$  it follows that  $T_R(x) = M + 1$ . We obtain that there exists  $\varepsilon > 0$  so that for each  $y \in J$  with  $\rho([x, y]_J) < \varepsilon$  either  $T_R(y) = M$  or  $T_R(y) = M + 1$ .

We conclude: there exists  $\varepsilon > 0$  so that for every  $y \in J$  with  $\rho([x, y]_J) < \varepsilon$  either  $T_R(x) = T_R(y)$  or  $|T_R(x) - T_R(y)| = 1$ . This completes the proof of the T-Jump Lemma.

Denote the length of a segment  $L \subset J$  by  $\rho(L)$ .

**GEOMETRIC LEMMA II.** There exists  $\delta > 0$  such that for every  $J$  in  $U(R)$ , and for each integer  $k \geq 1$ , the following holds:

- (1) Each component of  $A_k(J)$  contains components of  $D_k(J)$  and  $A_{k+1}(J)$ ;
- (2) Let  $A$  be an arbitrarily chosen component of  $A_k(J)$ . For each component  $D$  of  $D_k(J) \cap A$ , one has  $\rho(D)/\rho(A) \geq \delta$ , and each component  $U$  of  $A_{k+1}(J) \cap A$ , satisfies  $\rho(U)/\rho(A) \geq \delta$ .

**Proof of Geometric Lemma II.** Let  $J \in U(R)$ . Proof of (1). For  $k = 1$ , the assumptions (A1)-(A4) imply that  $A_1(J) = J$  contains at least two components of  $D_1(J)$ , and it contains at least one component of  $A_2(J)$ . Now assume  $k > 1$ , and let  $A$  be a component of  $A_k(J)$ . By the definition of  $A_k(J)$  and the assumptions on  $F$ , we have  $F^{k-1}(A)$  is a component of  $U(R)$ . Hence,  $A$  contains at least two components of  $D_k(J)$  and at least one component of  $A_{k+1}(J)$ .

**Proof of (2).** Let  $U$  be a neighborhood of  $\text{Inv}(R)$  on which a  $C^{1+\alpha}$  stable foliation  $\mathcal{F}^s$  exists, for some  $\alpha > 0$ . The case that a basic set is nontrivial periodic is similar to that of a nonperiodic basic set but the notation is more complicated. Therefore, we assume that every basic set in  $\text{Inv}(R)$  is either nontrivial nonperiodic or trivial. For each nontrivial nonperiodic basic set  $\Gamma$  let  $I^\alpha$  and the regions  $R_i(\Gamma)$ ,  $1 \leq i \leq N(\Gamma)$ , be as in Proposition 4.2, and let  $U_\Gamma$  be an open neighborhood of  $\Gamma$  such that (1)  $\bigcup_{i=1}^{N(\Gamma)} R_i(\Gamma) \subset U_\Gamma \subset U$ , (2) the set  $\tau^{-1}(I^\alpha \cap U_\Gamma)$  and its closure consist both of  $N(\Gamma)$  components, and (3) the map  $\varphi$  in Proposition 4.4 may be extended to  $\tau^{-1}(I^\alpha \cap U_\Gamma)$ . For each trivial basic set  $\Gamma$ , let  $U_\Gamma$  be an open neighborhood of  $\Gamma$  in  $U$  such that  $U_\Gamma \cap U_\Lambda$  is empty, for each basic set  $\Lambda$  in  $\text{Inv}(R) \setminus \Gamma$ . Select an integer  $K \geq 1$  such that the union of the  $U_\Gamma$ 's include  $A_K(J)$ ; the existence of  $K$  is guaranteed by the fact that  $A_k(J) \rightarrow W^s(\text{Inv}(R)) \cap J$  as  $k \rightarrow \infty$ . From the assumptions on  $F$  we obtain that the number of components of both  $A_k(J)$  and  $D_k(J)$  is finite for all  $k$ . For

every  $k \geq 1$  and each basic set  $\Gamma$  in  $\text{Inv}(R)$ , we define  $\delta_k(J) = \min_A \min_V \rho(V)/\rho(A)$  and  $\delta_k(J; \Gamma) = \min_A \{\min_V \rho(V)/\rho(A) : A \cap \Gamma \text{ is nonempty}\}$ , where the minimum is taken over all components  $V$  of the sets  $D_k(J)$  and  $A_{k+1}(J)$ , and all components  $A$  of the set  $A_k(J)$  such that  $V \subset A$ . Obviously,  $\delta_k(J) \leq \delta_k(J; \Gamma)$ , for all  $k$ .

Let  $\Gamma$  be a basic set in  $\text{Inv}(R)$ . Write  $U_\Gamma(R) = \{J \in U(R) : J \cap \Gamma \text{ is nonempty}\}$ . We first show: there exists  $\delta_\Gamma > 0$  such that for each  $J \in U_\Gamma(R)$ , for all  $k \geq 1$ , and for every component  $A$  of  $A_k(J)$  that intersects  $\Gamma$ , one has every component  $D$  of  $D_k(J) \cap A$  satisfies  $\rho(D)/\rho(A) \geq \delta_\Gamma$ , and each component  $U$  of  $A_{k+1}(J) \cap A$  satisfies  $\rho(U)/\rho(A) \geq \delta_\Gamma$ . The case that  $\Gamma$  is a periodic orbit is left to the reader. We assume that  $\Gamma$  is a nontrivial nonperiodic basic set.

Applying Proposition 4.4 and Geometric Lemma I we obtain that there exists  $\delta_n(J; \Gamma) > 0$  such that  $\delta_k(J; \Gamma) \geq \delta_n(J; \Gamma)$  for all  $k > K$ . We write  $\delta_\beta(J; \Gamma) = \min_{1 \leq k \leq K} \delta_k(J; \Gamma)$ , then  $\delta_k(J; \Gamma) \geq \delta_\beta(J; \Gamma)$  for all  $1 \leq k \leq K$ . Now we define  $\delta(J; \Gamma) = \min \{\delta_n(J; \Gamma), \delta_\beta(J; \Gamma)\} > 0$  and get  $\rho(V)/\rho(A) \geq \delta(J; \Gamma)$  for every component  $A$  of  $A_k(J)$  and every component  $V$  with  $V \subset A$ , where  $V$  is either in  $D_k(J)$  or in  $A_{k+1}(J)$ . Now, we define  $\delta_\Gamma = \inf \{\delta(J; \Gamma) : J \in U_\Gamma(R)\}$ . Since  $U_\Gamma(R)$  is compact we obtain  $\delta_\Gamma = \min \{\delta(J; \Gamma) : J \in U_\Gamma(R)\} > 0$ . Finally, since  $\Gamma$  was arbitrarily given, we define  $\delta = \min \{\delta_\Gamma : \Gamma \text{ basic set in } \text{Inv}(R)\}$ , and conclude  $\delta_k(J) \geq \delta$  for all  $k \geq 1$ . This completes the proof of Geometric Lemma II.

*Proof of Proposition 1. Let  $L$  be as in the proposition.*

'(i)  $\Rightarrow$  (ii)': We assume that there exists  $\varepsilon > 0$  such that every  $\varepsilon$ -refinement of  $\{a, b\}$  includes a PIM triple. If the interior of  $L$  does not include a point of  $\text{Inv}(R)$ , then  $A_\infty(J) \cap L$  is empty, and thus no  $\varepsilon$ -refinement of  $\{a, b\}$  includes a PIM triple. Hence, the interior of  $L$  contains a boundary point of  $D_k(J)$  for some integer  $k \geq 0$ . Therefore, the interior of  $L$  intersects  $A_\infty(J)$ .

'(ii)  $\Rightarrow$  (i)': Now we assume that the interior of  $L$  contains a point  $q$  of  $A_\infty(J) \cap \Gamma$ , for some basic set  $\Gamma$  in  $\text{Inv}(R)$ . Select integer  $M \geq 1$ , such that  $L$  contains a component  $A$  of  $A_M(J)$  that includes  $q$ . Let  $\delta > 0$  be as in the Geometric Lemma II. Now we select  $\varepsilon = \delta^2 \cdot \rho(A)$ . From the Geometric Lemma II we know that  $A$  contains at least two components of  $D_M(J)$  whose length of each of them is at least  $\delta \cdot \rho(A)$ , and  $A$  contains one or more components of  $A_{M+1}$  whose length of each of them is at least  $\delta \cdot \rho(A)$ . We obtain that each  $\varepsilon$ -refinement of  $a$  and  $b$  includes a PIM triple in  $A$ . This completes the proof of Proposition 1.

From now on, we fix  $\delta$  as in Geometric Lemma II and  $\varepsilon = \delta^2$ .

*Proof of Proposition 2. Let  $(a, c, b)$  be an Interior Maximum triple in  $J$ . First, we assume that  $T_R(a) \leq T_R(b) < T_R(c)$ .*

*Case 1. Assume  $k = \min_{a \leq x \leq b} T_R(x) < T_R(a)$ . Let  $D$  be the component of  $D_k(J)$  containing at least one point of  $[a, b]_J$ , for which  $T_R(y) = k$  for all  $y$  in  $D$ . Then  $D \subset \text{int}([a, b]_J) \subset A$ , where  $A$  is the component of  $A_k(J)$  for which  $D \subset A$ . Since  $\rho([a, b]_J) \leq \rho(A)$ , applying the Geometric Lemma II gives  $\rho(D)/\rho([a, b]_J) \geq \delta$ . Then, for every  $\beta$ -refinement  $P_\beta$  of  $(a, c, b)$ , with  $0 < \beta \leq \delta$  we have  $P_\beta \cap D \neq \emptyset$ . We obtain: for each  $p \in P_\beta \cap D$  either  $(p, c, b)$  or  $(a, c, p)$  is a PIM triple in  $P_\beta$ .*

*Case 2.* Assume  $\min_{a \leq x \leq b} T_R(x) \geq T_R(a)$  and  $T_R(c) \geq T_R(a) + 2 = m + 1$ . Then, by the *T-Jump Lemma*, there exists a component  $D$  of  $D_m(J)$  in the interval  $[a, c]_J$ . Since  $[a, b]_J$  lies in a component  $A$  of  $A_{m-1}(J)$ , the *Geometric Lemma II* implies  $\rho(D)/\rho([a, b]_J) \geq \delta$ . Hence, every  $\beta$ -refinement of  $(a, c, b)$  includes a point  $p$  of  $D$ , so  $(p, c, b)$  is a PIM triple, where  $0 < \beta \leq \delta$ .

*Case 3.* Assume  $T_R(c) = T_R(a) + 1 = m$  and that Case 1 does not apply. This implies  $T_R(b) = T_R(a)$ . Set  $\beta = \delta^2$ ; let  $P_\beta$  be a  $\beta$ -refinement of  $(a, c, b)$ , say  $P_\beta = \{x_i : 0 \leq i \leq N(\beta)\} \subset J$  with  $a = x_0 < x_1 < \dots < x_{N(\beta)} = b$  and  $x_k = c$  for some  $1 \leq k \leq N(\beta) - 1$ .

From the *Geometric Lemma II* we get that  $[a, b]_J$  contains a component  $D$  of  $D_{m+1}(J)$ , and  $\rho(D)/\rho([a, b]_J) \geq \delta^2$ . We obtain that every  $\beta$ -refinement of  $(a, c, b)$  includes a PIM triple for each  $0 < \beta \leq \delta^2$ .

The case  $T_R(b) \leq T_R(a) < T_R(c)$  is similar. The conclusion is: For  $\varepsilon = \delta^2$  we have: every  $\varepsilon$ -refinement of a PIM triple in  $J$  includes a PIM triple. This completes the proof of Proposition 2.

*Proof of Proposition 3. Left to the reader.*

Before we will prove Proposition 4, we will present a monotonicity property for the escape time map as well as an auxiliary observability result for Accessible PIM triple sequences.

**MONOTONICITY LEMMA.** *Let  $a$  and  $c$  be two points on  $J$ , and let  $P \subset [a, c]_J$  be a  $\beta$ -refinement of  $a$  and  $c$ , say  $P = \{x_i : 0 \leq i \leq N(\beta)\}$  and  $a = x_0 < x_1 < \dots < x_{N(\beta)} = c$ , where  $\beta > 0$ . Assume that  $T_R$  is monotonic on  $P$  (that is,  $T_R(x_{k+1}) \geq T_R(x_k)$ ,  $0 \leq k \leq N(\beta) - 1$ ), and  $T_R(c) > T_R(a)$ . Write  $m = \min \{T_R(x) : x \in [a, c]_J\}$ .*

*Then, for every  $\beta$ ,  $0 < \beta < \delta$ ,  $D_m(J) \cap [a, c]_J$  consists of one component, and it includes  $c$ .*

*Proof of the Monotonicity Lemma.* Let  $\beta$ ,  $a$ ,  $c$ ,  $P$ ,  $T_R$ , and  $m$  be as in the Lemma. By the definition of  $m$ , we know that  $[a, c]_J$  is contained in a component  $A$  of  $A_m(J)$ . Assume that  $0 < \beta < \delta$ .

Suppose that  $T_R(a) > m$ . Then there is a component  $D$  of  $D_m(J)$  such that  $D \subset [a, c]_J$ . (Note that neither  $a$  nor  $c$  is contained in  $D$ .) From *Geometric Lemma II* we know that  $\rho(D)/\rho([a, c]_J) \geq \rho(D)/\rho(A) \geq \delta > \beta$ ; this implies  $P \cap D \neq \emptyset$ . But this contradicts the assumption  $T_R$  is monotonic on  $P$ . Hence,  $m = T_R(a)$ .

Suppose  $D_m(J) \cap [a, c]_J$  consists of two components, say  $D$  and  $D'$ . We will assume  $D'$  includes  $a$ . The *Geometric Lemma II* implies there exists a component  $U$  of  $A_{m+1}(J)$  between  $D$  and  $D'$  such that  $\rho(U)/\rho([a, c]_J) \geq \rho(U)/\rho(A) \geq \delta > \beta$ . We obtain that  $P$  includes a PIM triple  $(a, c', b')$  with  $c' \in P \cap U$  and  $b' \in P \cap D$  (since both  $\rho(D)/\rho([a, c]_J) > \beta$  and  $\rho(U)/\rho([a, c]_J) > \beta$ ), which contradicts the monotonicity of  $T_R$  on  $P$ . This completes the proof of the Monotonicity Lemma.

**OBSERVABILITY LEMMA.** *Let  $P \subset J$  be an  $\varepsilon/3$ -refinement of an Interior Maximum triple  $(a_0, c_0, b_0)$  in  $J$ , and assume  $T_R(x_i) \geq T_R(a_0)$  for every  $x_i \in P$ . Let  $(a_0, c_1, b_1)$  be the PIM triple in  $P$ , in which  $b_1$  and  $c_1$  are selected as in the Accessible PIM triple procedure, and let  $a_1^0$  and  $a_1^1$  be defined as in the Accessible PIM triple procedure.*

If  $P$  is an  $\varepsilon$ -refinement of  $(a_0, c_1, b_1)$ , then

- (i) If  $a_1^0 > a_0$  then  $[a_0, a_1^0]_J$  does not intersect  $S(R)$ ; otherwise,
- (ii) if  $a_1^0 = a_0$  then  $T_R(b_1) > T_R(a_0)$ ,  $a_1^1 < c_1$  and  $[a_0, a_1^1]_J$  does not intersect  $S(R)$ .

*Proof of Observability Lemma.* Let  $P, (a_0, c_1, b_1)$ ,  $a_1^0$  and  $a_1^1$  be as in the Lemma, and assume  $P \cap [a_0, b_1]_J$  is an  $\varepsilon$ -refinement of  $(a_0, c_1, b_1)$ . Note that from this latter assumption it follows that  $P \cap [a_0, c_1]_J$  is a  $\beta$ -refinement of  $\{a_0, c_1\}$  for some  $0 < \beta < \delta$ .

Let  $m = \min \{T_R(x) : x \in [a_0, b_1]_J\}$ . The assumptions ' $T_R(x_i) \geq T_R(a_0)$  for all  $x_i \in P$ ', ' $P \cap [a_0, b_1]_J$  is an  $\varepsilon$ -refinement of  $(a_0, c_1, b_1)$ ', and the Geometric Lemma II imply that  $m = T_R(a_0)$ .

*Proof of (i).* Assume that  $a_1^0 > a_0$ . By the Monotonicity Lemma we obtain that  $T_R(x) = T_R(a_0)$  for all  $x \in [a_0, a_1^0]_J$ ; hence,  $[a_0, a_1^0]_J$  does not intersect  $S(R)$ .

*Proof of (ii).* Assume that  $a_1^0 = a_0$ . Suppose  $T_R(b_1) = T_R(a_0) = m$ . From the Geometric Lemma II and the assumptions we get that the interval  $[a_0, b_1]_J$  contains one component  $A$  of  $A_{m+1}(J)$ , and  $\rho(A)/\rho([a_0, b_1]_J) > \delta$ . Applying the Geometric Lemma II again, we get that there are at least 2 components  $U_1$  and  $U_2$  of  $D_{m+1}(J)$  and at least one component  $U_3$  of  $A_{m+2}(J)$  in  $A$ , and for each  $k$ ,  $1 \leq k \leq 3$ ,  $\rho(U_k)/\rho([a_0, b_1]_J) = (\rho(U_k)/\rho(A))(\rho(A)/\rho([a_0, b_1]_J)) > \delta^2 = \varepsilon$ . Hence, each  $U_k$ ,  $1 \leq k \leq 3$ , contains at least one point of  $P$ . This implies  $b_1$  is not the leftmost point in  $P$  that is the right point in a PIM triple, which contradicts the assumption. Conclusion:  $T_R(b_1) > T_R(a_0)$ .

The facts ' $(a_0, c_1, b_1)$  is a PIM triple' and ' $T_R(a_0) < T_R(b_1)$ ' imply  $T_R(c_1) \geq T_R(a_0) + 2$ . We obtain from the Geometric Lemma II that there is a component  $D$  of  $D_{m+1}(J)$  in  $[a_0, c_1]_J$  such that  $\rho(D)/\rho([a_0, b_1]_J) \geq \delta$ . Using the  $T$ -Jump Lemma, we obtain that there is a point  $q \in D \cap P$  with  $T_R(q) = T_R(a_0) + 1$  and for all  $x$  in  $P$  between  $a_0$  and  $q$  one has  $T_R(a_0) \leq T_R(x) \leq T_R(a_0) + 1$ . It follows that the point  $a_1^1$  exists. Applying the Monotonicity Lemma we obtain  $m = T_R(a_0) \leq T_R(x) \leq T_R(a_1^1) = m + 1$  for all  $x \in [a_0, a_1^1]_J$ ; hence,  $[a_0, a_1^1]_J$  does not intersect  $S(R)$ . This completes the proof of the Observability Lemma.

*Proof of Proposition 4.* Let  $\varepsilon$  be as in Proposition 2, and let  $\{(a_n, c_n, b_n)\}_{n \geq 0}$  be an Accessible PIM triple sequence in  $J$ , that is,  $(a_0, c_0, b_0)$  is an Interior Maximum triple and for  $n \geq 1$ ,  $(a_n, c_n, b_n)$  is obtained by the Accessible PIM triple procedure. For  $n \geq 0$ , let  $P_n$  be an  $\varepsilon/3$ -refinement of the Interior Maximum triple  $(a_n, c_n, b_n)$ , and recall that  $M_n = \min \{T_R(x_i) : x_i \in P_n, x_i < c_{n+1}\}$ . Further, we write  $m_n = \min \{T_R(x_i) : x_i \in P_n\}$ . Note that the Geometric Lemma II implies  $m_n = \min \{T_R(x) : x \in [a_n, b_n]_J\}$ .

We will show that there exists an integer  $N \geq 0$  such that for every integer  $n \geq N$ :  $T_R(a_n) = M_n$ ;  $|T_R(a_{n+1}) - T_R(a_n)| \leq 1$ ; and  $[a_n, a_{n+1}]_J$  does not intersect  $S(R)$ .

From the  $T$ -Jump Lemma, the Geometric Lemma II, and the assumption that  $\{(a_n, c_n, b_n)\}_{n \geq 1}$  is obtained by the Accessible PIM triple procedure we obtain for each  $n \geq 0$ , the following properties:

- (1) if  $T_R(a_n) > M_n$  then  $T_R(a_{n+1}) = M_n$ ;
- (2) if  $T_R(b_n) = m_n$  then  $T_R(b_{n+1}) \geq T_R(b_n) + 1$ ;

(3) if  $T_R(b_n) = m_n$  and  $M_n > m_n$  then  $m_{n+1} \geq m_n + 1$ ;

(4) if  $T_R(b_n) > m_n$  and  $M_n > m_n$  then  $T_R(b_{n+1}) \geq m_n$ .

These properties imply that there exists a minimal integer  $N \geq 0$  such that  $T_R(x_i) \geq M_N = m_N = T_R(a_N)$  for each  $x_i \in P_N$ .

**Case 1.**  $P_N$  is no  $\varepsilon$ -refinement of  $(a_N, c_{N+1}, b_{N+1})$ . Since  $a_{N+1} = a_N$ , we have (1)  $T_R(x_i) \geq M_{N+1} = m_{N+1} = T_R(a_{N+1})$  for each  $x_i \in P_{N+1}$ , and (2)  $[a_N, a_{N+1}]_J$  does not intersect  $S(R)$ . Obviously,  $T_R(x) = T_R(a_N)$  for all  $x$  in  $[a_N, a_{N+1}]_J$ .

**Case 2.**  $P_N$  is an  $\varepsilon$ -refinement of  $(a_N, c_{N+1}, b_{N+1})$ . First, assume that  $a_{N+1}^0 > a_N$ . By the Monotonicity Lemma, and the Observability Lemma we obtain for  $a_{N+1} = a_{N+1}^0$ : (1)  $T_R(x) = T_R(a_N)$  for all  $x \in [a_N, a_{N+1}]_J$ ,

(2)  $T_R(x_i) \geq M_{N+1} = m_{N+1} = T_R(a_{N+1})$  for each  $x_i \in P_{N+1}$ , and

(3)  $[a_N, a_{N+1}]_J$  does not intersect  $S(R)$ .

Now assume that  $a_{N+1}^0 = a_N$ . Applying the Monotonicity Lemma, and the Observability Lemma yields for  $a_{N+1} = a_{N+1}^1$ : (1)  $T_R(x) = T_R(a_N)$  for every  $x \in [a_N, a_{N+1}]_J$ , (2)  $T_R(x_i) \geq M_{N+1} = m_{N+1} = T_R(a_{N+1}) = T_R(a_N) + 1$  for each  $x_i \in P_{N+1}$ , and (3)  $[a_N, a_{N+1}]_J$  does not intersect  $S(R)$ .

By induction, one obtains the desired result. This completes the proof of Proposition 4.

*Proof of Proposition 5.* Left to the reader.

### 5. Discussion of the numerical procedures

Now we will return to the 'dynamic' question addressed in the beginning, namely, how can you numerically follow a trajectory on an invariant set for an arbitrarily long period of time?

A line segment  $[a, b]$  straddles the stable manifold of a point  $P$  if  $[a, b]$  intersects this manifold transversally. In the cases we study,  $P$  will be replaced by chaotic saddles (nontrivial basic sets) and  $[a, b]$  will straddle a subset of  $S(R)$ . Furthermore, in practice  $[a, b]$  will be very short and will be extremely close to the invariant set  $\text{Inv}(R)$ .

The numerical procedure goes as follows: (1) Choose (with some experimenting) a straight line segment  $I$ ; (2) Apply PIM triple procedure (refine and choose PIM triple interval) repeatedly until the length of the PIM triple interval is less than some distance  $\sigma$  (e.g.  $\sigma = 10^{-8}$ ); call this interval  $I_0 = \text{PIM}_\sigma(I)$ ; (3) For a straight line segment  $L$  with end points  $a$  and  $b$ , we write  $\text{PIM}_\sigma(L)$  to denote either  $[a, b]$  if  $|[a, b]| < \sigma$  or the resulting interval when some PIM triple procedure is applied until an interval of length less than  $\sigma$  is reached. Note that this operator depends only on the end points of  $L$ . The basic process then is iterating  $\text{PIM}_\sigma(F(L))$ . While  $F(L)$  is an interval, only  $F(a)$  and  $F(b)$  are relevant. Thus we obtain  $I_{n+1} = \text{PIM}_\sigma(F(I_n))$ , a sequence of straight line segments.

We thus obtain a trajectory of tiny straight line segments  $I_n$  and to the precision of the computer (about  $10^{-14}$ ) we typically have  $I_{n+1} \subset F(I_n)$ , and selecting any point  $x_n$  from  $I_n$ , perhaps the midpoint, we have that  $|x_{n+1} - F(x_n)|$  is small, typically of the order of  $\sigma$ . Since computers can never be expected to produce true trajectories

(except in trivial cases such as fixed points), we may say  $\{x_n\}_{n \geq 0}$  is a numerical trajectory. We call the sequence of intervals  $\{I_n\}_{n \geq 0}$  a *saddle straddle trajectory* because the interval straddles a piece of the stable set  $S(R)$  of a chaotic saddle set. It typically approximates (after a few iterates) a basic set in the invariant set (which is a chaotic saddle) in the interesting cases. Furthermore, a saddle trajectory approximates the trajectory of a point in the Static Restraint Problem. Despite the complexity of the construction, we will refer to  $x_{n+1}$  as the 'iterate' of  $x_n$ .

*Remark.* In practice we find that every  $\varepsilon$ -refinement of two points  $\{a, b\}$ , with  $\varepsilon = 1/30$ , includes several PIM triples. In computing the sequence of PIM triples  $(a_n, c_n, b_n)$  defined by the Accessible PIM triple procedure, once either case 3(iii) or 3(iv) holds, and if  $c$  is more than  $\varepsilon \cdot |b - a|$  from  $a$  and  $b$ , then it can be shown that every  $\varepsilon$ -refinement of the end points  $\{a, b\}$  of a PIM triple  $(a, c, b)$  includes a PIM triple; in the computer program we do not use  $c$  at all. For the examples in this paper and in [NY] we find that the Accessible PIM triple procedure leads to accessible fixed points or periodic points, which is in agreement with the fact that all the accessible points for two dimensional hyperbolic systems are on the stable manifolds of finitely many periodic points.

In this paper we have shown that our procedures are valid in ideal situations. We find it works well in practice even in less than ideal cases. From the examples in [NY], we have seen that the PIM triple procedure works quite well for a variety of dynamical systems.

It is important to ask if such straddle trajectories represent true trajectories of the system. In other words, does there exist a true trajectory of the system that shadows (i.e., stays close to) the numerical trajectory obtained by the PIM triple procedure? When a map is sufficiently hyperbolic on the invariant set in question, Bowen [B] obtained a result saying that each noisy trajectory in the nonwandering set can be shadowed by a true trajectory if the perturbation is small; see [B] for the precise statement. We will say that  $\text{Inv}(R)$  satisfies the 'no cycle condition' if for every family of basic sets  $\Gamma_{k(1)}, \dots, \Gamma_{k(M)}$  in  $\text{Inv}(R)$  such that the stable set of  $\Gamma_{k(i)}$  has a nonempty intersection with the unstable set of  $\Gamma_{k(i+1)}$  for all  $1 \leq i < M$ , the stable manifold of  $\Gamma_{k(M)}$  does not intersect the unstable manifold of  $\Gamma_{k(1)}$ . Assuming  $\text{Inv}(R)$  satisfies the 'no cycle condition' and  $\delta$  is sufficiently small, we can show that every saddle straddle trajectory of a two dimensional uniformly hyperbolic system with a chaotic saddle obtained by the PIM triple procedure, can be shadowed by a true trajectory for as long as the saddle straddle trajectory can be computed.

## 6. Concluding remarks

6.1. *Higher dimensional systems.* One of the ingredients in the analysis of the validity of the PIM triple procedures in dimension two, is the existence of a  $C^{1+\alpha}$  foliation  $\mathcal{F}$  on a neighborhood of a basic set. The existence of such a foliation for the two dimensional case, is guaranteed by a result due to De Melo [M]. Unfortunately, the existence of a foliation  $\mathcal{F}$  on a neighborhood of a basic set in higher dimensions is not known, see e.g. [PT].

Let from now on, the dimension  $n \geq 3$ . Let  $F$  be an Axiom A diffeomorphism, let  $R$  be a saddle-hyperbolic transient region for which  $\dim E^u = 1$ , and assume that for each basic set  $\Gamma$  in  $\text{Inv}(R)$  there exists a  $C^{1+\alpha}$  stable foliation  $\mathcal{F}^s$  on a neighborhood of  $\Gamma$ , for some  $\alpha > 0$ . Then the Propositions 1, 2, 3, 4, and 5 are still valid. The proof is almost the same, except instead of Propositions 4.1 and 4.2 one should use the properties of Markov partitions of basic sets, see Bowen [B].

**6.2. Order of Differentiability of the Diffeomorphism.** We assumed that the diffeomorphism  $F$  is  $C^3$ . This assumption implied the existence of a  $C^{1+\alpha}$  expanding map, for some  $\alpha > 0$ , in Proposition 4.4. If  $F$  is of class  $C^2$ , then it is known that such an expanding map is  $C^1$ . We would like to point out, that the Hölder exponent  $\alpha$  is only used to obtain (2) in the proof of the Geometric Lemma I. Fortunately, we can prove the Geometric Lemma I (in particular property (2)) for the  $C^1$ -map  $\varphi$  of Proposition 4.4 by combining the techniques of the proof of Proposition 6 in [Ne] and Lemma 5.5 in [Nu]. Thus in fact, it is sufficient to assume  $F$  is  $C^2$  to guarantee the main results of the paper.

#### REFERENCES

- [AY] K. T. Alligood & J. A. Yorke. Accessible saddles on fractal basin boundaries. Preprint 1989.
- [B] R. Bowen. Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms. *Lecture Notes in Mathematics* 470, Springer Verlag: Berlin, 1975
- [BR] R. Bowen & D. Ruelle. The ergodic theory of Axiom A flows. *Invent. Math.* 29 (1975), 181-202.
- [DN] R. Devaney & Z. Nitecki. Shift automorphisms in the Hénon mapping. *Commun. Math. Phys.* 67 (1979), 137-146.
- [GOY] C. Grebogi, E. Ott & J. A. Yorke. Basin boundary metamorphoses: changes in accessible boundary orbits. *Physica* 24D (1987), 243-262.
- [GNOY] C. Grebogi, H. E. Nusse, E. Ott & J. A. Yorke. Basic sets: sets determine the dimension of basin boundaries. In: *Dynamical Systems*, ed. J. C. Alexander. *Proceedings of the University of Maryland 1986-87. Lecture Notes in Math.* 1342, pp. 220-250 Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, 1988.
- [GH] J. Guckenheimer & P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Applied Mathematical Sciences* 42, Springer Verlag: New York, 1983
- [M] W. de Melo. Structural stability of diffeomorphisms on two-manifolds. *Invent. Math.* 21 (1973), 233-246.
- [NP] S. Newhouse & J. Palis. Hyperbolic nonwandering sets on two-dimensional manifolds. In *Dynamical Systems*, pp. 293-301, ed. M. M. Peixoto. Academic Press: New York and London, 1973.
- [Ne] S. E. Newhouse. The abundance of wild hyperbolic sets and non-smooth stable sets for diffeomorphisms. *Publ. Math. I.H.E.S.* 50 (1979), 101-151.
- [Ni] Z. Nitecki. *Differentiable Dynamics*, MIT Press, Cambridge, 1971
- [Nu] H. E. Nusse. Asymptotically periodic behaviour in the dynamics of chaotic mappings. *SIAM J. Appl. Math.* 47 (1987), 498-515.
- [NY] H. E. Nusse & J. A. Yorke. A procedure for finding numerical trajectories on chaotic saddles. *Physica* D36 (1989), 137-156.
- [PT] J. Palis & F. Takens. Homoclinic bifurcations and hyperbolic dynamics. *16º Colóquio Brasileiro Matemática*, IMPA, 1987.
- [S] S. Smale. Differentiable dynamical systems. *Bull. Amer. Math. Soc.* 73 (1967), 747-817.
- [Y] J. A. Yorke. *DYNAMICS. A Program for IBM PC Clones*. 1987, 1988.

## Embedology

Tim Sauer,<sup>1</sup> James A. Yorke,<sup>2</sup> and Martin Casdagli<sup>3</sup>

Received May 10, 1991

Mathematical formulations of the embedding methods commonly used for the reconstruction of attractors from data series are discussed. Embedding theorems, based on previous work by H. Whitney and F. Takens, are established for compact subsets  $A$  of Euclidean space  $R^n$ . If  $n$  is an integer larger than twice the box-counting dimension of  $A$ , then almost every map from  $R^n$  to  $R^n$ , in the sense of prevalence, is one-to-one on  $A$ , and moreover is an embedding on smooth manifolds contained within  $A$ . If  $A$  is a chaotic attractor of a typical dynamical system, then the same is true for almost every *delay-coordinate map* from  $R^n$  to  $R^n$ . These results are extended in two other directions. Similar results are proved in the more general case of reconstructions which use moving averages of delay coordinates. Second, information is given on the self-intersection set that exists when  $n$  is less than or equal to twice the box-counting dimension of  $A$ .

**KEY WORDS:** Embedding; chaotic attractor; attractor reconstruction; probability one; prevalence; box-counting dimension; delay coordinates

### 1. INTRODUCTION

In this work we give theoretical justification of data embedding techniques used by experimentalists to reconstruct dynamical information from time series. We focus on cases in which trajectories of the system under study are asymptotic to a compact attractor. We state conditions that ensure that the map from the attractor into reconstruction space is an embedding, meaning that it is one-to-one and preserves differential information. Our approach integrates and expands on previous results on embedding by Whitney<sup>(29)</sup> and Takens.<sup>(27)</sup>

<sup>1</sup> Department of Mathematical Sciences, George Mason University, Fairfax, Virginia 22030.

<sup>2</sup> Institute of Physical Science and Technology, University of Maryland, College Park, Maryland 20742.

<sup>3</sup> Santa Fe Institute, Santa Fe, New Mexico 87501. Current address: Tech Partners, 4 Stamford Forum, 8th Floor, Stamford, Connecticut 06901.

Whitney showed that a generic smooth map  $F$  from a  $d$ -dimensional smooth compact manifold  $M$  to  $R^{2d+1}$  is actually a diffeomorphism on  $M$ . That is,  $M$  and  $F(M)$  are diffeomorphic. We generalize this in two ways: first, by replacing "generic" with "probability-one" (in a prescribed sense), and second, by replacing the manifold  $M$  by a compact invariant set  $A$  contained in  $R^k$  that may have noninteger box-counting dimension (boxdim). In that case, we show that almost every smooth map from a neighborhood of  $A$  to  $R^n$  is one-to-one as long as

$$n > 2 \text{ boxdim}(A)$$

We also show that almost every smooth map is an embedding on compact subsets of smooth manifolds within  $A$ . This suggests that embedding techniques can be used to compute positive Lyapunov exponents (but not necessarily negative Lyapunov exponents). The positive Lyapunov exponents are usually carried by smooth unstable manifolds on attractors. We give precise definitions of one-to-one, embedding, and almost every in the next section.

Takens dealt with a restricted class of maps called delay-coordinate maps. A delay-coordinate map is constructed from a time series of a single observed quantity from an experiment. Because of this, a typical delay-coordinate map is much more likely to be accessible to an experimentalist than a typical map. Takens<sup>(27)</sup> showed that if the dynamical system and the observed quantity are generic, then the delay-coordinate map from a  $d$ -dimensional smooth compact manifold  $M$  to  $R^{2d+1}$  is a diffeomorphism on  $M$ .

Our results generalize those of Takens<sup>(27)</sup> in the same two ways as for Whitney's theorem. Namely, we replace generic with probability-one and the manifold  $M$  by a possibly fractal set. Thus, for a compact invariant subset  $A$  of  $R^k$ , under mild conditions on the dynamical system, almost every delay-coordinate map  $F$  from  $R^k$  to  $R^n$  is one-to-one on  $A$  provided that  $n > 2 \text{ boxdim}(A)$ . Also, any manifold structure within  $A$  will be preserved in  $F(A)$ . These results hold for lower box-counting dimension (see Section 4) if boxdim does not exist. The ambient space  $R^k$  can be replaced by a  $k$ -dimensional smooth manifold in the general case. In addition, we have made explicit the hypotheses on the dynamical system (discrete or continuous) that are needed to ensure that the delay-coordinate map gives an embedding. In particular, only  $C^1$  smoothness is needed. For flows, the delay must be chosen so that there are no periodic orbits whose period is exactly equal to the time delay used or twice the delay. (A finite number of periodic orbits of a flow whose periods are  $p$  times the delay are allowed for  $p \geq 3$ .) Further, we explain what happens

case that  $n \leq 2 \cdot \text{boxdim}(A)$ . In that case we put bounds on the on of the self-intersection set, which is the set on which the one-property fails. Finally, we give more general versions of the delay-coordinate theorem which deals with filtered delay coordinates, which are more versatile and useful applications of embedding methods.

There are no analogues of these results where the box-counting dimension is replaced by Hausdorff dimension (see Theorem 4.7 and the section that follows). In an Appendix to this work written by I. Kan, sets are described of compact subsets of  $R^k$ , for any positive integer  $k$ , which have Hausdorff dimension  $d = 0$ , and which are difficult to project one-to-one. The requirement  $n > 2d$  discussed above translates in this case to  $n > 0$ . However, every projection of such a set to  $R^n$ ,  $n < k$ , fails to be one-to-one.

In Section 2 we explain the new version of the Whitney and Takens embedding theorems. In Section 3 we discuss filtered delay coordinates. Section 4 contains proofs of the results.

## WHY TO EMBED MANIFOLDS AND FRACTAL SETS

### Fractal Whitney Embedding Prevalence Theorem

Assume  $\phi$  is a flow on Euclidean space  $R^k$ , generated, for example, by a dynamical system of  $k$  differential equations. Assume further that all trajectories are asymptotic to an attractor  $A$ . The study of long-time behavior of the system will involve the study of the set  $A$ .

In a typical scientific experiment, the phase space  $R^k$  cannot be directly seen. The experimenter tries to infer properties of the system by measurements. Since each state of the dynamical system is uniquely identified by a point in phase space, a measured quantity is a function from phase space to the real number line. If  $n$  independent quantities  $Q_1, \dots, Q_n$  are measured simultaneously, then with each point in phase space is associated a point in Euclidean space  $R^n$ . We can then talk about the measurement function

$$F(\text{state}) = (Q_1, \dots, Q_n)$$

maps  $R^k$  to  $R^n$ .

For example, suppose all trajectories in phase space  $R^k$  are attracted to a periodic cycle. Thus,  $A$  is topologically a circle lying in  $R^k$ . Imagine that two available measurement quantities  $Q_1$  and  $Q_2$  are plotted in the plane. Then there is a measurement map  $F$  from  $A$  to  $R^2$  given by  $F(x) = (Q_1, Q_2)$ . To what extent are the properties of the hidden attractor  $A$  preserved in the observable "reconstruction space"  $R^2$ ?

The answer depends on how the circle is mapped to  $R^2$  under  $F$ . Consider the case where  $R^k = R^3$  and  $Q_1$  and  $Q_2$  are simply the two coordinate functions  $x_1$  and  $x_2$ . In Fig. 1a, the relative position of the points is preserved upon projection, and we may view  $F(A)$  as a faithful reconstruction of the attractor  $A$ . If distinct points on the attractor  $A$  map under  $F$  to distinct points on  $F(A)$ , we say that  $F$  is *one-to-one* on  $A$ .

In the case of Fig. 1b, on the other hand, two different states of the dynamical system have been identified together in  $F(A)$ . In the reconstruction space, which is all the experimenter actually sees, the two distinct states cannot be distinguished, and information has been lost.

The one-to-one property is useful because the state of a deterministic dynamical system, and thus its future evolution, is completely specified by a point in phase space. Suppose that at a given state  $x$  one observes the value  $F(x)$  in the reconstruction space, and that this is followed 1 sec later by a particular event. If  $F$  is one-to-one, each appearance of the measurements represented by  $F(x)$  will be followed 1 sec later by the same event. This is because there is a one-to-one correspondence between the attractor points in phase space and their images in reconstruction space. There is predictive power in finding a one-to-one map.

Perhaps the measurements  $F(x)$  would not be repeated precisely. Yet if the map  $F$  is reasonable, similar measurements will predict similar events. This approach to prediction and noise reduction of data is being pursued by a number of research groups.

Although most of the interest lies in the case that  $A$  is an attractor of a dynamical system, the main question can be posed in more generality. Let  $A$  be a compact subset of Euclidean space  $R^k$ , and let  $F$  map  $R^k$  to another Euclidean space  $R^n$ . Under what conditions can we be assured that  $A$  is "embedded" in  $R^n$  by typical maps  $F$ ?

The precise definition of embedding involves differential structure. A one-to-one map is a map that does not collapse points, that is, no two points are mapped to the same image point. An embedding is a map that does not collapse points or tangent directions. Thus, to define embedding, we need to be working on a compact set  $A$  that has well-defined tangent spaces.

Let  $A$  be a compact smooth differentiable manifold. (Here, as in the remainder of the paper, the word *smooth* will be used to mean continuously differentiable, or  $C^1$ .) A smooth map  $F$  on  $A$  is an *immersion* if the derivative map  $DF(x)$  (represented by the Jacobian matrix of  $F$  at  $x$ ) is one-to-one at every point  $x$  of  $A$ . Since  $DF(x)$  is a linear map, this is equivalent to  $DF(x)$  having full rank on the tangent space. This can happen whether or not  $F$  is one-to-one. Under an immersion, no differential structure is lost in going from  $A$  to  $F(A)$ .

An *embedding* of  $A$  is a smooth diffeomorphism from  $A$  onto its image  $F(A)$ , that is, a smooth one-to-one map which has a smooth inverse. For a compact manifold  $A$ , the map  $F$  is an embedding if and only if  $F$  is a one-to-one immersion. Figure 1a shows an example of an embedding of a circle into the plane. Figure 1b shows an immersion that is not one-to-one, and Fig. 1c shows a one-to-one map that fails to be an immersion.

Whether or not a typical map from  $A$  to  $R^n$  is an embedding of  $A$  depends on the set  $A$ , and on what we mean by "typical." A celebrated result of this type is the embedding genericity theorem of Whitney,<sup>(29)</sup> which says that if  $A$  is a smooth manifold of dimension  $d$ , then the set of maps into  $R^{d+1}$  that are embeddings of  $A$  is an open and dense set in the  $C^1$ -topology of maps.

The fact that the set of embeddings is *open* in the set of smooth maps means that given each embedding, arbitrarily small perturbations will still be embeddings. The fact that the set of embeddings is *dense* in the set of maps means that every smooth map, whether it is an embedding or not, is arbitrarily near an embedding. One would like to conclude from Whitney's

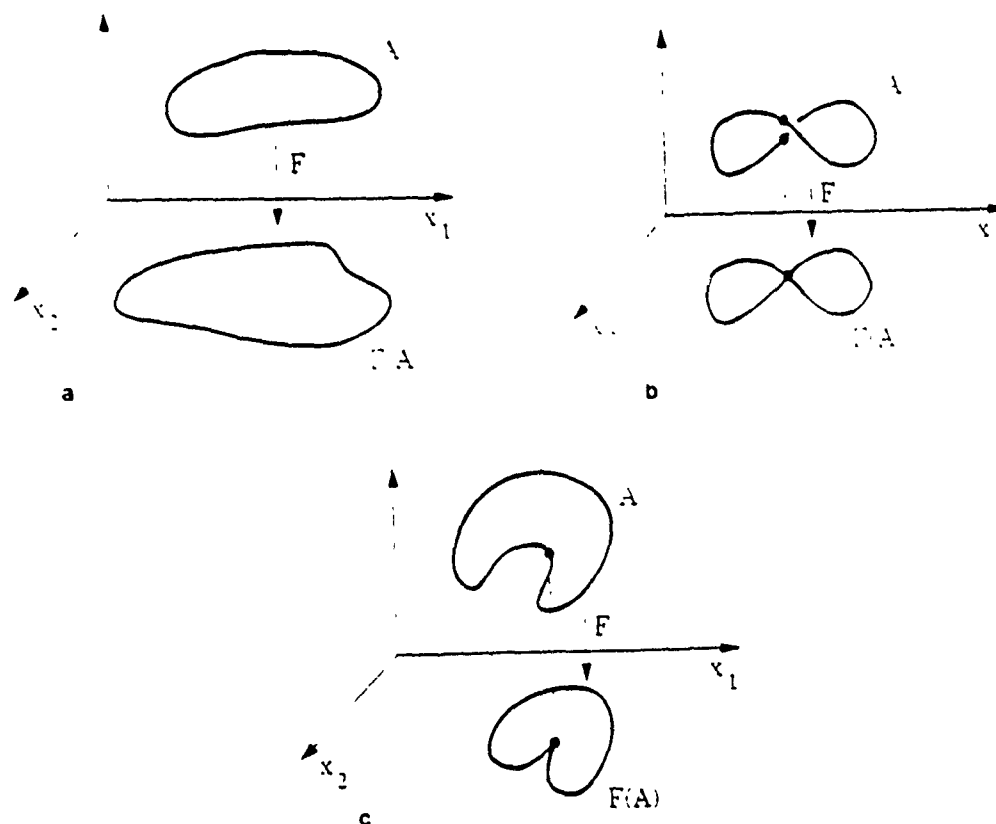


Fig. 1 (a) An embedding  $F$  of the smooth manifold  $A$  into  $R^2$  (b) An immersion that fails to be one-to-one. (c) A one-to-one map that fails to be an immersion.

theorem that  $n = 2d + 1$  simultaneous measurements are typically sufficient to reconstruct a  $d$ -dimensional state manifold  $A$  in the measurement space  $R^n$ .

However, open dense subsets, even of Euclidean space, can be thin in terms of probability. There are standard examples, many from recent studies in dynamics, of open dense sets that have arbitrarily small Lebesgue measure, and therefore arbitrarily small probability of being realized.

A well-known example is the phenomenon of Arnold tongues. Consider the family of circle diffeomorphisms

$$g_{\omega,k}(x) = x + \omega + k \sin x \pmod{2\pi}$$

where  $0 \leq \omega \leq 2\pi$  and  $0 \leq k < 1$  are parameters. For each  $k$  we can define the set

$$\text{Stab}(k) = \{0 \leq \omega \leq 2\pi : g_{\omega,k} \text{ has a stable periodic orbit}\}$$

For  $0 < k < 1$ , the set  $\text{Stab}(k)$  is a countable union of disjoint open intervals of positive length, and is an open dense subset of  $[0, 2\pi]$ . However, the total length (Lebesgue measure) of the open dense set  $\text{Stab}(k)$  approaches zero as  $k \rightarrow 0$ . For small  $k$ , the probability of landing in this open dense set is very small. See ref. 3 for more details.

With such examples in mind, an experimentalist would like to make a stronger statement than that the set of embeddings is an open and dense set of smooth maps. Instead, one would like to know that the particular map that results from analyzing the experimental data is an embedding with probability one.

The problem with such a statement is that the space of all smooth maps is infinite-dimensional. The notion of probability one on infinite-dimensional spaces does not have an obvious generalization from finite-dimensional spaces. There is no measure on a Banach space that corresponds to Lebesgue measure on finite-dimensional subspaces. Nonetheless, we would like to make sense of "almost every" map having some property, such as being an embedding. Following ref. 24, we propose the following definition of prevalence.

**Definition 2.1.** A Borel subset  $S$  of a normed linear space  $V$  is *prevalent* if there is a finite-dimensional subspace  $E$  of  $V$  such that for each  $v$  in  $V$ ,  $v + e$  belongs to  $S$  for (Lebesgue) almost every  $e$  in  $E$ .

We give the distinguished subspace  $E$  the nickname of *probe space*. The fact that  $S$  is prevalent means that if we start at any point in the ambient space  $V$  and explore along the finite-dimensional space of directions specified by  $E$ , then almost every point encountered will lie in  $S$ .

Notice that any space containing a probe space for  $S$  is itself a probe space for  $S$ . In other words, if  $E'$  is any finite-dimensional space containing  $E$ , then perturbations of any element of  $V$  by elements of  $E'$  will be in  $S$  with probability one. This is a simple consequence of the Fubini theorem.<sup>(22)</sup>

From this fact it is easy to see that a prevalent subset of a finite-dimensional vector space is simply a set whose complement has zero measure. Also, the union or intersection of a finite number of prevalent sets is again prevalent. We will often use the notion of prevalence to describe subsets of functions. It follows from the definition that prevalent implies dense in the  $C^k$ -topology for any  $k$ . More generally, prevalent implies dense in any normed linear space.

When a condition holds for a prevalent set of functions, it is usually illuminating to determine the smallest, or most efficient, probe subspace  $E$ . This corresponds to the minimal amount of perturbation that must be available to the system in order for the condition to hold with virtual certainty.

As stated above, for subsets of finite-dimensional spaces the term prevalent is synonymous with "almost every," in the sense of outside a set of measure zero. When there is no possibility of confusion, we will say that "almost every" map satisfies a property when the set of such maps is prevalent, even in the infinite-dimensional case. For example, consider convergent Fourier series in one variable, which form an infinite-dimensional vector space with basis  $\{e^{inx}\}_{n=-\infty}^{\infty}$ . In the sense of prevalence, almost every Fourier series has nonzero integral on  $[0, 2\pi]$ . The probe space  $E$  in this case is the one-dimensional space of constant functions. If  $E'$  is a vector space of Fourier series which contains the constant functions, then for every Fourier series  $f$ , the integral of  $f + e$  will be nonzero for almost every  $e$  in  $E'$ .

With this definition, we introduce a prevalence version of the Whitney embedding theorem.

**Theorem 2.2 (Whitney Embedding Prevalence Theorem).** Let  $A$  be a compact smooth manifold of dimension  $d$  contained in  $R^k$ . Almost every smooth map  $R^k \rightarrow R^{2d+1}$  is an embedding of  $A$ .

In particular, given any smooth map  $F$ , not only are there maps arbitrarily near  $F$  that are embeddings, but in the sense of prevalence, *almost all* of the maps near  $F$  are embeddings. The probe space  $E$  of Definition 2.1 is the  $k(2d+1)$ -dimensional space of linear maps from  $R^k$  to  $R^{2d+1}$ . This theorem, which is proved in Section 4, gives a strengthening of the traditional statement of the Whitney embedding theorem.

It is quite interesting that Whitney later proved the different result that under the same circumstances, there *exists* an embedding into  $R^{2d}$ . (This

could be called the Whitney embedding existence theorem.) However, an existence theorem is of little help to an experimentalist, who needs information about maps near the particular one that happens to be available. Knowledge that an embedding exists sheds little information on the particular  $F$  under study.

The example of Fig. 1b shows that the dimension  $2d+1$  of Theorem 2.2 is the best possible. The map  $F$  is not one-to-one on the twisted circle  $A$ , thus does not embed  $A$  into  $R^2$ . Further, no nearby map (even in the  $C^\infty$ -topology) embeds  $A$ . On the other hand, if a given map of the circle  $A$  into  $R^3$  was not one-to-one, there would necessarily be a prevalent set of nearby maps that are embeddings.

The first main goal of this section was to express Whitney's embedding theorem (and Takens' theorem; see below) in this probabilistic sense. The second is to extend Whitney's theorem to sets  $A$  that are not manifolds. Here we use the fractal dimension known as box-counting dimension.

The box-counting (or capacity) dimension of a compact set  $A$  in  $R^n$  is defined as follows. For a positive number  $\varepsilon$ , let  $A_\varepsilon$  be the set of all points within  $\varepsilon$  of  $A$ , i.e.,  $A_\varepsilon = \{x \in R^n : |x - a| \leq \varepsilon \text{ for some } a \in A\}$ . Let  $\text{vol}(A_\varepsilon)$  denote the  $n$ -dimensional outer volume of  $A_\varepsilon$ . Then the *box-counting dimension* of  $A$  is

$$\text{boxdim}(A) = n - \lim_{\varepsilon \rightarrow 0} \frac{\log \text{vol}(A_\varepsilon)}{\log \varepsilon}$$

if the limit exists. If not, the upper (respectively, lower) box-counting dimension can be defined by replacing the limit by the  $\liminf$  (resp.,  $\limsup$ ). When the box-counting dimension exists, the approximate scaling law

$$\text{vol}(A_\varepsilon) \approx \varepsilon^{n-d}$$

holds, where  $d = \text{boxdim}(A)$ .

There are several equivalent definitions of box-counting dimension. For example,  $R^n$  can be divided into  $\varepsilon$ -cubes by a grid based, say, at points whose coordinates are  $\varepsilon$ -multiples of integers. Let  $N(\varepsilon)$  be the number of boxes that intersect  $A$ . Then

$$\text{boxdim}(A) = \lim_{\varepsilon \rightarrow 0} \frac{\log N(\varepsilon)}{-\log \varepsilon}$$

with similar provisions for upper and lower box-counting dimension. The scaling in this case is

$$N(\varepsilon) \approx \varepsilon^{-d}$$

Even if we know the box-counting dimension of an attractor  $A$ , Theorem 2.2 gives no estimate on the lowest dimension for which almost every map is an embedding. Suppose we know that  $A$  is the invariant set of a flow on  $R^{100}$ , and that the box-counting dimension of  $A$  is 1.4. In the absence of any knowledge about the containment of  $A$  in a smooth manifold of dimension less than 100, the use of Theorem 2.2 to get a one-to-one reconstruction requires the use of maps into  $R^{201}$ . In fact, the smallest smooth manifold that contains the 1.4-dimensional attractor may indeed have dimension 100. But as the next result shows, one can do much better: almost every reconstruction map into  $R^3$  will be one-to-one on  $A$ .

**Theorem 2.3** (Fractal Whitney Embedding Prevalence Theorem). Let  $A$  be a compact subset of  $R^k$  of box-counting dimension  $d$ , and let  $n$  be an integer greater than  $2d$ . For almost every smooth map  $F: R^k \rightarrow R^n$ ,

1.  $F$  is one-to-one on  $A$
2.  $F$  is an immersion on each compact subset  $C$  of a smooth manifold contained in  $A$ .

The proof of the one-to-one half of the fractal Whitney embedding prevalence theorem may be sketched as follows. Choose any bounded finite-dimensional space  $E$  of smooth maps  $F$  so that varying  $F$  by elements of  $E$  results in perturbing  $F(x) - F(y)$  throughout  $R^n$  for each pair  $x \neq y$  in  $A$ . For example, the probe space  $E$  can be taken to be the space of linear maps from  $R^k$  to  $R^n$ . Then the probability (measured in  $E$ ) that the perturbed  $F(x)$  and  $F(y)$  lie within  $\epsilon$  is on the order of  $\epsilon^n$ . Similarly, if  $B_1$  and  $B_2$  are  $\epsilon$ -boxes on  $A$ , the probability that  $F(B_1)$  and  $F(B_2)$  intersect is on the order of  $\epsilon^n$ . Here we assume that there is a bound on the magnification of  $F$ , as when  $F$  is a smooth map near the compact set  $A$ . The set  $A$  can be covered by essentially  $\epsilon^{-d}$  boxes of size  $\epsilon$ , and the number of pairs of boxes is proportional to  $(\epsilon^{-d})^2$ . The probability that no distinct pair of boxes collide in the image  $F(A)$  is proportional to  $(\epsilon^{-d})^2 \epsilon^n = \epsilon^{n-2d}$ . If  $n > 2d$ , this probability of choosing a perturbation of  $F$  that fails to be one-to-one is negligible for small  $\epsilon$ . More precise details of the proof, as well as the immersion part, are in Section 4.

## 2.2. Fractal Delay Embedding Prevalence Theorem

Despite the beauty of Whitney's embedding theorem, it is rare for a scientist to be able to measure a large number of independent quantities simultaneously. In fact, it is a rather subtle problem to decide whether two different simultaneous measurements are indeed independent. These problems can be sidestepped to some degree by introducing the use of

*delay coordinates.* In this approach, only one measurable quantity is needed.

In a typical experiment, the single measurable quantity is sampled at intervals  $T$  time units apart. The resulting list of samples  $\{Q_i\}$  is called a *time series*. Think of the measurable quantity as an observation function  $h$  on the state space  $R^k$  on which the dynamical system  $\phi$  is acting. Each reading  $Q_i = h(x_i)$  is the result of evaluating the observation function  $h$  at the current state  $x_i$ .

**Definition 2.4.** If  $\phi$  is a flow on a manifold  $M$ ,  $T$  is a positive number (called the *delay*), and  $h: M \rightarrow R$  is a smooth function, define the *delay-coordinate map*  $F(h, \phi, T): M \rightarrow R^n$  by

$$F(h, \phi, T)(x) = (h(x), h(\phi_{-T}(x)), h(\phi_{-2T}(x)), \dots, h(\phi_{-(n-1)T}(x)))$$

To start with a simple example, let  $A$  be a periodic orbit of the flow  $\phi$ . We found above that in the absence of dynamics, three independent coordinates are required to embed  $A$  in reconstruction space, or more precisely, that almost every smooth map  $F = (f_1, f_2, f_3)$  from a neighborhood of  $A$  to  $R^3$  is an embedding on  $A$ .

Now the situation is different. Instead of three functions  $f_1, f_2, f_3$  that must be independent, there is a single function  $h$ , and the corresponding map  $F(h, \phi, T)$  pictured in Fig. 2. We want to know that for almost every function  $h$  from  $A$  to the real numbers  $R$ , the delay-coordinate map  $F(h, \phi, T)$  from  $A$  into  $R^n$  is an embedding. It should be stressed that this does not follow from Theorems 2.2 and 2.3. The maps  $F(h, \phi, T)$  form a restricted subset of all maps; whether they contain enough variation to perturb away self-crossings of  $A$  needs to be determined. In fact, the general

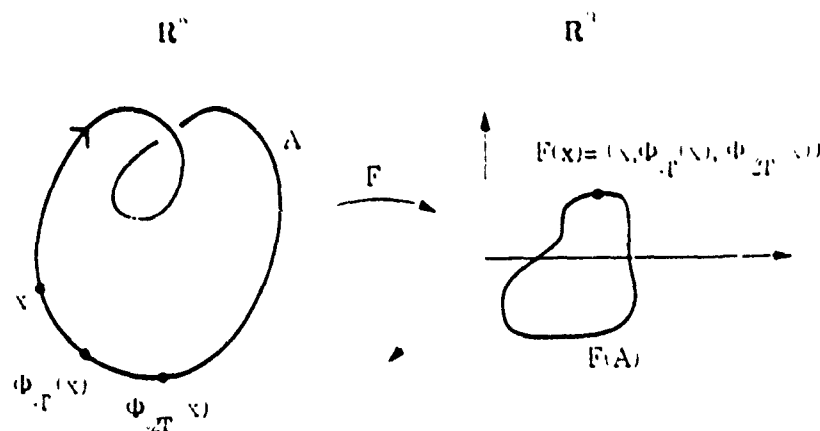


Fig. 2. The attractor on the left is mapped using delay coordinates into the reconstruction space on the right.

answer is that they do not contain enough variation. Extra hypotheses on the dynamical system  $\phi$  are required to ensure that almost every  $h$  leads to an embedding of  $A$ .

To see the need for extra hypotheses, consider the case that  $A$  is a periodic orbit of a continuous dynamical system whose period is equal to the sampling interval  $T$ . Topologically,  $A$  is a circle. In this case,  $F(h, \phi, T)$  cannot be one-to-one for *any* observation function  $h$ . Let  $x$  be a point on the topological circle  $A$ . Since the period of  $A$  is  $T$ ,  $h(x) = h(\phi_{-T}(x)) = \dots = h(\phi_{-(n-1)T}(x))$ , so that  $F = F(h, \phi, T)$  maps  $x$  to the diagonal line  $\{(x_1, \dots, x_n); x_1 = \dots = x_n\}$  in  $R^n$ . A circle cannot be mapped continuously to a line (in this case, the diagonal line in  $R^n$ ) in a one-to-one fashion. See Fig. 3.

The one-to-one property also fails when  $A$  is a periodic orbit of period  $2T$ . Define the function  $d(x) = h(x) - h(\phi_{-T}(x))$  on  $A$ . The function  $d$  is either identically zero or it is nonzero for some  $x$  on  $A$ , in which case it has the opposite sign at the image point  $\phi_{-T}(x)$ , and changes sign on  $A$ . In any case,  $d(x)$  has a root  $x_0$  on  $A$ . Since the period of  $A$  is  $2T$ , we have  $h(x_0) = h(\phi_{-T}(x_0)) = h(\phi_{-2T}(x_0)) = \dots$ . Then  $F(h, \phi, T)$  maps  $x_0$  and  $\phi_{-T}(x_0)$  to the same point in  $R^n$ . If  $x_0$  and  $\phi_{-T}(x_0)$  are distinct, this says that  $F$  is not one-to-one. If  $x_0 = \phi_{-T}(x_0)$ , then the orbit actually has period  $T$ , and  $F$  fails to be one-to-one as above. In the presence of periodic orbits of period  $2T$ ,  $F(h, \phi, T)$  cannot be one-to-one for any observation function  $h$ .

On the other hand, when  $A$  is a periodic orbit of period  $3T$ , or any period not equal to  $T$  or  $2T$ , there is no such problem. In this case the delay-coordinate map of a periodic orbit  $A$  into  $R^n$  is an embedding for almost every observation function  $h$  as long as the reconstruction dimension is at least three. The statement for more general attractors  $A$  is as follows.

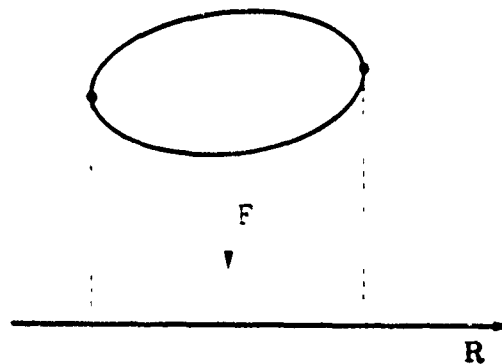


Fig. 3. A two-to-one map from a topological circle to the real line.

**Theorem 2.5 (Fractal Delay Embedding Prevalence Theorem).** Let  $\Phi$  be a flow on an open subset  $U$  of  $R^n$ , and let  $A$  be a compact subset of  $U$  of box-counting dimension  $d$ . Let  $n > 2d$  be an integer, and let  $T > 0$ . Assume that  $A$  contains at most a finite number of equilibria, no periodic orbits of  $\Phi$  of period  $T$  or  $2T$ , at most finitely many periodic orbits of period  $3T, 4T, \dots, nT$ , and that the linearizations of those periodic orbits have distinct eigenvalues. Then for almost every smooth function  $h$  on  $U$ , the delay coordinate map  $F(h, \Phi, T): U \rightarrow R^n$  is:

1. One-to-one on  $A$ .
2. An immersion on each compact subset  $C$  of a smooth manifold contained in  $A$ .

Where Takens<sup>[27]</sup> showed that the delay-coordinate maps generically (in the  $C^2$ -topology) give embeddings of smooth manifolds of dimension  $d$ , we substitute compact sets of box-counting dimension  $d$ , and replace generic with prevalent.

The assumption of Theorem 2.5 that there are no periodic orbits of period  $T$  or  $2T$  can be satisfied by choosing the time delay  $T$  to be sufficiently small. In fact, if we assume that the vector field on  $A$  satisfies a Lipschitz condition, that is,  $\dot{x} = V(x)$ , where  $\|V(x) - V(y)\| \leq L\|x - y\|$ , then it is known<sup>[10]</sup> that each periodic orbit must have period at least  $2\pi/L$ . Hence, if  $T < \pi/L$ , there will be no periodic orbits of period  $T$  or  $2T$ .

Theorem 2.5 assumes  $n > 2d$  to avoid self-intersection of the reconstructed image of  $A$ . To see that this requirement cannot be relaxed in general, consider the case  $d=1$ ,  $n=2d=2$  shown in Fig. 4a. Let the observation function  $h$  be the coordinate function  $x_1$ , and consider the delay coordinate map  $R^k \rightarrow R^2$  defined by

$$F(x_1, \Phi, T) = (x_1(x), x_1(\Phi_{-T}(x)))$$

In the situation illustrated in Fig. 4a,  $x_1(\Phi_{-T}(b)) < x_1(\Phi_{-T}(a)) < x_1(a) = x_1(b)$ , and  $x_1(\Phi_{-T}(c)) < x_1(\Phi_{-T}(d)) < x_1(c) = x_1(d)$ . Setting  $F = F(x_1, \Phi, T)$ , this means that in the reconstruction space  $R^2$ ,  $F(a)$  lies directly above  $F(b)$ , and  $F(d)$  lies directly above  $F(c)$ . See Fig. 4b. The map  $F$  is continuous on the trajectory, so there is a continuous path, parametrized by  $x_1$ , connecting  $F(a)$  and  $F(c)$ . There is also such a path connecting  $F(b)$  and  $F(d)$ . According to Fig. 4b, there must be a value of  $x_1$  in between where the curves meet, and two different points on the circle map together under  $F$ . Otherwise said, somewhere in between there is an  $x_1$  coordinate such that the upper and lower parts of the trajectory advance the same amount in the  $x_1$  direction during the time interval  $T$ , and thus have identical delay coordinates. The map  $F(h, \Phi, T)$  is not an embedding.

If the observation function or flow is perturbed a small amount, the same topological argument can be made. Thus, this example is robust. No small perturbation of the map is an embedding.

Theorem 2.5 is a special case of a statement about diffeomorphisms. Before stating that version, we redefine delay coordinate maps for diffeomorphisms.

**Definition 2.6.** If  $g$  is a diffeomorphism of an open subset  $U$  of  $R^n$ , and  $h: U \rightarrow R$  is a function, define the *delay coordinate map*  $F(h, g): U \rightarrow R^n$  by

$$F(h, g)x = (h(x), h(g(x)), h(g^2(x)), \dots, h(g^{n-1}(x)))$$

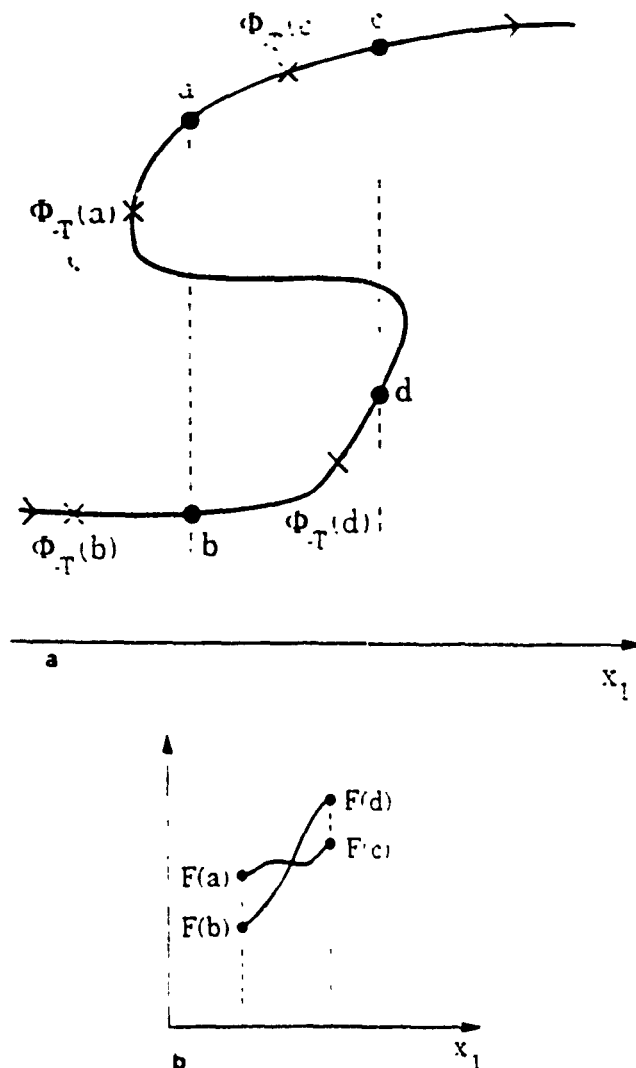


Fig. 4 (a) A trajectory of a flow that cannot be mapped using two delay coordinates in a one-to-one way. (b) The point at which the paths cross corresponds to a set of delay coordinates shared by two points on the trajectory.

We get the previous theorem by substituting  $g = \phi_{-T}$  in the following statement.

**Theorem 2.7.** Let  $g$  be a diffeomorphism on an open subset  $U$  of  $R^d$ , and let  $A$  be a compact subset of  $U$ ,  $\text{boxdim}(A) = d$ , and let  $n > 2d$  be an integer. Assume that for every positive integer  $p \leq n$ , the set  $A_p$  of periodic points of period  $p$  satisfies  $\text{boxdim}(A_p) < p - 2$ , and that the linearization  $Dg^p$  for each of these orbits has distinct eigenvalues.

Then for almost every smooth function  $h$  on  $U$ , the delay coordinate map  $F(h, g): U \rightarrow R^n$  is:

1. One-to-one on  $A$ .
2. An immersion on each compact subset  $C$  of a smooth manifold contained in  $A$ .

**Remark 2.8.** The probe space for this prevalent set can be taken to be any set  $h_1, \dots, h_n$  of polynomials in  $k$  variables which includes all polynomials of total degree up to  $2n$ . Given any smooth function  $h_0$  on  $U$ , for almost all choices of  $x = (x_1, \dots, x_k)$  from  $R^k$ , the function  $\tilde{h}_i = h_0 + \sum_{k=1}^n x_k h_k$  satisfies properties 1 and 2.

**Remark 2.9.** The proof of Theorem 2.7 is easily extended to the more general case where the reconstruction map  $F$  consists of a mixture of lagged observations. The more general result says that

$$F(x) = (h_1(x), \dots, h_1(g^{n_1-1}(x)), \dots, h_p(x), \dots, h_p(g^{n_p-1}(x)))$$

satisfies the conclusions of Theorem 2.7 as long as  $n_1 + \dots + n_p \geq 2d$  and the corresponding conditions on the periodic points are satisfied. Those conditions are that  $\text{boxdim}(A_p) < p - 2$  for  $p = \max\{n_1, \dots, n_p\}$ .

The reconstruction of chaotic trajectories using independent coordinates from a time series was advocated in 1980 by Packard *et al.*<sup>(21)</sup> The delay-coordinate map is attributed in that work to a communication with D. Ruelle. The method actually illustrated in ref. 21 is somewhat different: namely, it is to use the value  $u_t$  of the time series and its time derivatives  $\dot{u}_t, \ddot{u}_t, \dots$  as independent coordinates.

In 1981, Takens<sup>(27)</sup> published the first mathematical results on the delay-coordinate map. Around the same time, Roux and Swinney<sup>(28)</sup> exhibited plots of delay-coordinate reconstructions of experimental data from the Belousov-Zhabotinski reaction.

In 1985, Eckmann and Ruelle<sup>(9)</sup> took the idea one step further and suggested examining not only the delay coordinates of a point, but also the relation between the delay coordinates of a point and the next point which occurs  $T$  time units later. In principle, one can then approximate not only

the attractor, but the attractor together with its dynamics. Since ref. 9 it has become common practice to gather points that are close in reconstruction space, and use their next images to construct a low-order parametric model which approximates the dynamics in a small region. This idea has begun to be used for prediction and noise reduction applications. See, for example, refs. 1, 6, 12, 13, 15, 16, 18, and 28.

### 2.3. Self-Intersection

In the case that the reconstruction dimension  $n$  is not greater than twice the box-counting dimension  $d$  of the set  $A$ , the map  $F$  in the fractal Whitney embedding prevalence theorem (Theorem 2.3) will often not be an embedding. However, if  $d < n$ , most of  $A$  will still be embedded. In the case that  $A$  is a smooth manifold of dimension  $d$ , almost every  $F$  will be an embedding outside a subset of  $A$  of dimension at most  $2d - n$ . If  $d < n$ , then  $2d - n < d$ , and so this exceptional subset will have positive codimension in  $A$ .

If  $A$  is simply a compact set of box-counting dimension  $d$ , then the situation is slightly different. We will call the pair  $x, y$  of points  $\delta$ -distant if the distance between them is at least  $\delta$ . Then we define the  $\delta$ -distant self-intersection set of  $F$  to be the subset of  $A$  consisting of all  $x$  such that there is a  $\delta$ -distant point  $y$  with  $F(x) = F(y)$ ; that is,

$$\Sigma(F, \delta) = \{x \in A; F(x) = F(y) \text{ for some } y \in A, |x - y| \geq \delta\}$$

Then the result is that for every  $\delta > 0$ , the lower box-counting dimension of the  $\delta$ -distant self-intersection set  $\Sigma(F, \delta)$  is at most  $2d - n$  for almost every  $F$ . A precise statement is given by the next theorem.

**Theorem 2.10** (Self-Intersection Theorem). Let  $A$  be a compact subset of  $R^k$  of box-counting dimension  $d$ , let  $n \leq 2d$  be an integer, and let  $\delta > 0$ . For almost every smooth map  $F: R^k \rightarrow R^n$ :

1. The  $\delta$ -distant self-intersection set  $\Sigma(F, \delta)$  of  $F$  has lower box-counting dimension at most  $2d - n$ .
2.  $F$  is an immersion on each compact subset  $C$  of an  $m$ -manifold contained in  $A$  except on a subset of  $C$  of dimension at most  $2m - n - 1$ .

For example, consider mapping a circle to the real line. In this case  $d = m = n = 1$ , and Theorem 2.10 says that a prevalent set of  $F$  are immersions outside a zero-dimensional set. This is clear from Fig. 3, where the zero-dimensional set consists of a pair of points. The map is at least 2 to 1 outside this set, and hence nowhere an embedding.

On the other hand, setting  $d = m = 1$  and  $n = 2$  in the theorem, we see that a prevalent set of maps  $F$  from the circle to the plane are immersions, and are embeddings outside a zero-dimensional subset. Thus, the maps shown in Figs. 1a and 1b are of the prevalent type, immersions which are one-to-one except for at most a discrete (zero-dimensional) set of points. Figure 1c, on the other hand, is nonprevalent. Almost any map near  $F$  will perturb away the cusp.

There is also a self-intersection version of the fractal delay embedding prevalence theorem (Theorem 2.5) which one gets by making the obvious changes. Thus, if  $n \leq 2d$ , then for each  $\delta > 0$  there exists a subset  $\Sigma(F, \delta)$ , whose box-counting dimension is at most  $2d - n$ , on which the delay-coordinate map fails to be one-to-one. Note that the result is independent of  $\delta > 0$ . If  $M$  is a closed subset of an  $m$ -manifold contained in  $A$ , then there is a subset  $E_\delta$  of  $M$  of dimension at most  $2m - n - 1$  on which the map fails to be an immersion.

#### 2.4. How Many Delay Coordinates Do You Need?

When using a delay coordinate map (or filtered delay coordinate map, described in the next section) to examine the image  $F(A)$  in  $R^n$  of a set  $A$  in  $R^k$ , the choice of  $n$  depends on the objective of the investigation. Different choices of  $n$  suffice for the different goals of prediction, calculation of dimension and Lyapunov exponents, and the determination of the stability of periodic orbits.

To compute the dimension of  $A$ , all that is required is that

$$\dim F(A) = \dim A \quad (2.1)$$

whether the dimension being used is box-counting, Hausdorff, information, or correlation dimension. The latter two depend on a probability density on  $A$  and  $F(A)$ . It is shown in ref. 24 that for the case of Hausdorff dimension, the equality (2.1) holds for almost every measurable map  $F$ , in the sense of prevalence, as long as  $n \geq \dim(A)$ . The probe space of perturbations for this result is the space of all linear transformations from  $R^k$  to  $R^n$ . Mattila<sup>(19)</sup> proved that equality (2.1) holds for almost every orthogonal projection  $F$ .

It is somewhat surprising that there are examples for which (2.1) does not hold for any map  $F$  when box-counting dimension is used, even under the hypothesis  $n > \text{boxdim}(A)$ . An example of this type is given in ref. 25. However, in most cases of compact sets which arise in dynamical systems, we expect Hausdorff dimension to equal box-counting dimension.

In practical situations, if attempts to measure  $\text{boxdim}(A)$  result in answers dependent on  $n$ , where  $n > \text{boxdim}(A)$ , then the variation would

seem to be a numerical artifact, since there is no theoretical justification for which of the values of  $n$  greater than  $\text{boxdim}(A)$  gives the more accurate result. The usual technique is to increase  $n$  until the observed dimension of  $\text{boxdim } F(A)$  reaches a plateau, and to use this result. The resulting number might be called the *plateau dimension*. While the plateau dimension may indeed give the best numerical estimate of the dimension of  $A$ , there does not seem to be theoretical or numerical justification of this bias, and the question needs further investigation. Notice that  $n > \text{boxdim}(A)$  does not guarantee that almost every  $F$  is one-to-one, but that is not required for dimension calculation.

If the objective is to use  $F(A)$  to predict the future behavior of trajectories, then it is sufficient to have the map  $F$  be one-to-one, in which case  $n > 2 \cdot \text{boxdim}(A)$  is needed. Knowing the current state in  $F(A)$  is sufficient to predict the future of the trajectory (at least in the short run). In the situation of Fig. 1b, on the other hand, prediction on the periodic orbit  $A$  would still be possible, except when the trajectory was at the midpoint of the "figure eight."

If the objective is to compute the Lyapunov exponents of the system, it is necessary to ask which exponents are to be computed. For a simple example, suppose the attractor  $A$  is a periodic orbit. Then the best possible result of the examination of  $F(A)$  is to observe that 0 is a Lyapunov exponent. The other exponents, presumably all negative, cannot be observed without introducing perturbations. More generally, if an attractor  $A$  lies on a manifold of dimension  $m$  (as a 2.2-dimensional attractor might lie on a three-dimensional manifold), it will certainly be impossible to measure more than  $m$  true exponents from an embedding, even if the reconstructed image  $F(A)$  lies in  $R^n$  with  $n > m$ . There are no criteria for determining the smallest manifold containing  $A$ .

Theorems 2.3 and 2.5 say that if  $n > 2 \cdot \text{boxdim}(A)$ , then almost every  $F$  is an embedding of all smooth manifolds that lie in  $A$ . The smooth manifolds we have in mind are the surface corresponding to the unstable directions on the attractor  $A$ , that is, the unstable manifolds. Under an embedding, the differential information is preserved along smooth directions, such as unstable manifolds, indicating that positive Lyapunov exponents should be computable from the image  $F(A)$ .

The stable manifolds, on the other hand, will be likely to intersect  $A$  in a Cantor set. The image of a Cantor set in  $F(A)$  may be quite compressed. For example, a set which is the product of five Cantor sets whose dimensions sum to 0.5 might be mapped to a one-dimensional line in  $F(A)$ . It seems difficult to recover any exponents in these directions from knowledge of the reconstructed dynamics in  $F(A)$ .

The self-intersection results in Section 2.3 are aimed at another kind of

question. A relevant experiment involving a vibrating ribbon is described in refs. 8 and 26. In this case, the Poincaré map has an attractor whose dimension was experimentally calculated to be 1.2. The investigators were interested in determining the eigenvalues of the linearization of a period-3 point on the attractor.

Using a delay-coordinate map of the attractor into  $R^2$  did not result in a one-to-one map, which is consistent with our results in Section 2.2. Theorem 2.10 of Section 2.3, which deals with self-intersection, suggests that the subset  $\Sigma$  of  $A$  on which the map into  $R^2$  fails to be one-to-one should have dimension at most  $2d - n = 2 \times 1.2 - 2 = 0.4$ . They found that the self-intersection set looked like a finite set. If  $\Sigma$  indeed has dimension 0.4 or less, as we would expect, then the set  $\Sigma$  would be unlikely to include the periodic point in question, and the delay-coordinate map would be expected to be one-to-one in a neighborhood of that orbit. Numerical investigations of the dynamics near the periodic orbit revealed that the dynamics did appear to be two-dimensional, and the researchers were able to estimate numerically the eigenvalues of the orbit at these points.

### 3. THE DELAY COORDINATE MAP AND FILTERS

#### 3.1. Main Results

So far, we have defined the delay coordinate map  $x \rightarrow F(h, g)x$  from the hidden phase space  $R^k$  to the reconstruction space  $R^n$ . Under suitable conditions on the diffeomorphism  $g$ , the delay coordinate map  $F(h, g)$  is an embedding for almost all observation functions  $h$ . In this formulation, information from the previous  $n$  time steps is used to identify a state of the original dynamical system in  $R^k$ .

For purposes of measuring quantitative invariants of the dynamical systems, noise reduction, or prediction, it may be advantageous to create an embedding that identifies a state with information from a larger number of previous time steps. However, working with embeddings in  $R^n$  is difficult for large  $n$ . A way around this problem is to incorporate large numbers of previous data readings by "averaging" their contributions in some sense. This problem has also been treated in ref. 7.

To this end, generalize the delay-coordinate map  $F(h, g): R^k \rightarrow R^n$ ,

$$F(h, g)x = (h(x), h(g(x)), \dots, h(g^{n-1}(x)))^T$$

where the superscript  $T$  denotes transpose, by defining the *filtered delay-coordinate map*  $F(B, h, g): R^k \rightarrow R^n$  to be

$$F(B, h, g)x = BF(h, g)x \quad (3.1)$$

where  $B$  is an  $n \times w$  constant matrix. Thus, each coordinate of  $F(B, h, g)x$  is a linear combination of the  $w$  coordinates of  $F(h, g)x$ . Here we are considering the case where  $g$  is a diffeomorphism, for notational convenience. Everything we say applies to a flow  $\phi$  by setting  $g$  equal to the time  $-T$  map of the flow. We will call  $w$  the *window length* of the reconstruction, since there are  $w$  evenly-spaced observations used. We call  $n$  the *reconstruction dimension*, since  $R^n$  is the range space of the map. We may as well assume that  $n \leq w$  and that  $B$  has rank  $n$ ; otherwise we could throw away some rows of  $B$  without losing information. Assuming that  $B$  is a fixed matrix restricts the filter to be a linear multidimensional moving average (MA) filter. Autoregressive (AR) filters in general can change the dimension of the attractor.<sup>(14-20)</sup>

If  $B$  is the identity matrix (denoted  $I$ ), the map is the original Takens delay coordinate map. As stated in the previous section, in that case,  $F(I, h, g) = F(h, g)$  is almost always an embedding as long as  $n$  is greater than twice the box-counting dimension of the attractor and the periodic points of period  $p$  less than  $n$  have distinct eigenvalues and make up a set of  $\text{boxdim} < p/2$ .

Under filtering, some complications are caused by the existence of periodic cycles. On the other hand, the next theorem states that in the absence of cycles of length smaller than the window length  $w$ , every moving average filter  $B$  gives a faithful representation of the attractor.

**Theorem 3.1** (Filtered Delay Embedding Prevalence Theorem). Let  $U$  be an open subset of  $R^k$ ,  $g$  be a smooth diffeomorphism on  $U$ , and let  $A$  be a compact subset of  $U$ ,  $\text{boxdim}(A) = d$ . For a positive integer  $n > 2d$ , let  $B$  be an  $n \times w$  matrix of rank  $n$ . Assume  $g$  has no periodic points of period less than or equal to  $w$ . Then for almost every smooth function  $h$ , the delay coordinate map  $F(B, h, g): U \rightarrow R^n$  is,

1. One-to-one on  $A$ .
2. An immersion on each closed subset  $C$  of a smooth manifold contained in  $A$ .

The probe space for perturbing  $h$  can be taken to be any space of polynomials in  $k$  variables which includes all polynomials of total degree up to  $2w$ . Furthermore, in case  $n \leq 2d$ , the results of Theorem 3.1 hold outside exceptional subsets of  $A$  precisely as in Theorem 2.10.

For example, consider the  $3 \times 9$  matrix

$$B = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \quad (3.2)$$

Then

$$F(B, h, g)x = \left( \frac{1}{m} (h(x) + h(g(x)) + h(g^2(x)) + \dots + h(g^{m-1}(x))) \right)$$

Although the map  $F(B, h, g)$  uses information from 9 different lags, the "moving average" reconstruction space is only 3-dimensional. According to the theorem, if the dynamical system  $g$  has no periodic points of period less than  $m=9$ , then  $F(B, h, g)$  is an embedding for almost all observation functions  $h$ .

**Remark 3.2.** When the diffeomorphism  $g$  has periodic points, certain special choices of filters  $B$  will cause self-intersection to occur at the periodic points. However, under the genericity hypotheses on the dynamical system of Theorem 2.5, for example, almost all choices of an  $n \times m$  matrix  $B$  imply the conclusions of Theorem 3.1. This follows from Remarks 3.4 and 3.5. A more detailed view of the effect of periodic points of the dynamical system is given in Sections 3.3 and 3.4.

### 3.2. Examples of Filters

In this section we will list some examples of filters that may be useful in given situations. The easiest example is a simple averaging filter. For any integers  $m, n$ , let  $B$  be a  $n \times nm$  matrix of form

$$B = \begin{pmatrix} 1/m & \dots & 1/m & & & \\ & & & 1/m & \dots & 1/m \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & 1/m \dots 1/m \end{pmatrix} \quad (3.3)$$

where there are  $m$  nonzero entries in each row. In the presence of noise, this filter should perform well compared to the more standard delay-coordinate embedding which uses every  $m$ th reading and discards the rest.

A more sophisticated noise filter was suggested in ref. 5 for a slightly different purpose, and elaborated on in the very readable ref. 2, where it is used for dimension measurements. It is based on the singular value decomposition from matrix algebra, also known as principal component analysis. Let  $y_1, \dots, y_L$  be the reconstructed vectors in  $R^n$ , where  $L$  is the length of the

data series. Following Broomhead and King,<sup>(5)</sup> define the  $L \times w$  trajectory matrix

$$A = \frac{1}{\sqrt{L}} \begin{pmatrix} y_1^T \\ \vdots \\ y_L^T \end{pmatrix}$$

where the  $y_i^T$  are treated as row vectors. The covariance matrix of this multivariate distribution is  $A^T A$ . The off-diagonal entries of  $A^T A$  measure the statistical dependence of the variables.

The singular value decomposition<sup>(14)</sup> of the  $L \times w$  matrix  $A$ , where  $L \geq w$ , is

$$A = U^T S U^T \quad (3.4)$$

where  $U$  is an  $L \times L$  orthogonal matrix,  $U^T$  is a  $w \times w$  orthogonal matrix (this means that  $U^T U = I$ ,  $U^T U^T = I$ ), and  $S$  is an  $L \times w$  diagonal matrix (meaning that the entries  $\sigma_i$  of  $S$  are zero if  $i \neq j$ ). By rearranging the rows and columns of  $U$  and  $U^T$ , we can arrange for the singular values of  $A$  to satisfy  $\sigma_{11} \geq \sigma_{22} \geq \dots \geq \sigma_{ww} \geq 0$ . The bottom  $L - w$  rows of  $S$  are zero.

The singular value decomposition suggests the use of the filter  $B = U^T$ . That is, instead of plotting the vectors  $y_1, \dots, y_L$  in reconstruction space  $R^n$ , plot the vectors  $U^T y_1, \dots, U^T y_L$ . One immediate positive consequence of this change of variables is the statistical linear independence of the new variables. The covariance matrix of the new trajectory matrix

$$\frac{1}{\sqrt{L}} \begin{pmatrix} (U^T y_1)^T \\ \vdots \\ (U^T y_L)^T \end{pmatrix} = U^T$$

is  $(AU)^T AU = S^T S$ , a diagonal matrix.

In practice, one can do better than  $B = U^T$ . This is because some of the nonzero singular values are dominated by noise. A rule of thumb is to ignore (by setting to zero) all singular values below the noise floor of the experimental data. Ignoring all but the largest  $k$  singular values is equivalent to letting the filter  $B$  in Eq. (3.1) be the top  $k$  rows of  $U^T$ . The rows of  $U^T$  are orthogonal, so  $B$  is still full rank. Theorem 3.1 implies that  $F(B, h, g)$  will typically be one-to-one and immersive.

This program was followed in ref. 2, in the context of measuring the correlation dimension of chaotic attractors in a stable way. They used a filter  $B$  that consisted of the rows of  $U^T$  that corresponded to singular values above  $10^{-4}$ .

### 3.3. Conditions on Periodic Orbits Which Imply One-to-One

For special filters  $B$ , conclusions 1 and 2 of Theorem 3.1 can fail, but only for periodic points. That is, some periodic points of period less than  $w$  may be mapped together under the map  $F(B, h, g)$ .

For example, assume

$$B = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \quad (3.5)$$

and assume that  $g$  has a period-4 orbit, that is,  $g^4(x) = x$ . Then for any  $h$ ,  $F(B, h, g)$  maps all four points of the period-4 orbit to the same point in  $R^3$ , so  $F(B, h, g)$  fails to be one-to-one. There is no way for any observation function to distinguish the four points, since their outputs are being averaged over the entire cycle. Thus, the filtered delay coordinate map fails, for all observation functions  $h$ , to be one-to-one.

A similar problem occurs with the filter

$$B = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \quad (3.6)$$

Now

$$\begin{aligned} F(B, h, g)x &= \left( \frac{1}{2}(h(x) + h(g^2(x))), \right. \\ &\quad \left. \frac{1}{2}(h(g(x)) + h(g^3(x))), \right. \\ &\quad \left. \frac{1}{2}(h(g^2(x)) + h(g^4(x))) \right) \end{aligned}$$

Assume that the period-four orbit of  $g$  consists of  $x_0, x_1 = g(x_0)$ ,  $x_2 = g^2(x_0)$ , and  $x_3 = g^3(x_0)$ . Now  $x_0$  and  $x_2$  are mapped to the same point in the reconstruction space  $R^3$  by  $F(B, h, g)$ , and the same goes for  $x_1$  and  $x_3$ . Again, the map cannot be one-to-one for any  $h$ .

A second obvious problem can be illustrated when the dynamical system has more than one fixed point. No matter how  $h$  is chosen, the filter

$$B = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \quad (3.7)$$

maps all fixed points to the origin in  $R^3$ , violating the one-to-one condition.

In each of these situations, the underlying dynamical system  $g$  may dictate that some periodic points will become identified under a particular

filter  $B$ , no matter how "generic" the observation function  $h$ . On the other hand, these identifications occur only at periodic points. Further, even in the case of periodic points, it turns out that the restrictions on  $B$  exemplified by the three cases above are the *only* restrictions. That is, if these are avoided, then  $F(B, h, g)$  is one-to-one for a prevalent set of observation functions  $h$ .

To be more precise about these restrictions, we need to make some definitions. For each positive integer  $p$ , denote by  $A_p$  the set of period- $p$  points of  $g$  lying on  $A$ . That is,  $A_p = \{x \in A : g^p(x) = x\}$ . Let  $I_n$  denote the  $n \times n$  identity matrix and  $(\cdot, \cdot)$  denote greatest common divisor. We will use the convention that  $(p, 0) = 0$ . For integers  $p > q \geq 0$ , define the  $p \times (p - (p, q))$  matrix

$$C_{pq} = \begin{pmatrix} I_{p-(p,q)} & & \\ & \ddots & \\ -I_{(p,q)} & & -I_{(p,q)} \end{pmatrix} \quad (3.8)$$

Define  $C_{pq}^w$  to be the  $w \times (p - (p, q))$  matrix formed by repeating the block  $C_{pq}$  vertically, and for a positive integer  $n$ , define  $C_{pq}^n$  to be the matrix formed by the top  $n$  rows of  $C_{pq}^w$ .

**Theorem 3.3.** Let  $U$  be an open subset of  $R^d$ , let  $g$  be a smooth diffeomorphism on  $U$ , and let  $A$  be a compact subset of  $U$  of box-counting dimension  $d$ . Let  $w$  and  $n$  be integers satisfying  $w \geq n > 2d$ . Assume that  $B$  is an  $n \times w$  matrix of rank  $n$  which satisfies:

$$\text{A1. } \text{rank } BC_{pq}^w \geq 2 \cdot \text{boxdim}(A_p) \text{ for all } 1 \leq p \leq w.$$

$$\text{A2. } \text{rank } BC_{pq}^n \geq \text{boxdim}(A_p) \text{ for all } 1 \leq q < p \leq w.$$

Then for almost every smooth function  $h$ ,  $F(B, h, g)$  is one-to-one on  $A$ .

**Remark 3.4.** Note that  $\text{rank } C_{pq} = p - (p, q)$ , and so  $\text{rank } C_{pq}^n = \min\{n, p - (p, q)\}$ . It follows that  $\text{rank } BC_{p0}^w \geq \min\{n, p\}$  and  $\text{rank } BC_{pq}^n \geq \min\{n, p/2\}$  for  $B = I_n$ , and also for almost every  $n \times w$  matrix  $B$ .

To illustrate the restrictions that Theorem 3.3 puts on moving average filters, assume that  $B$  is the  $3 \times 6$  matrix (3.5). In particular, the filter  $B$  must satisfy condition A2 for  $p = 4$ ,  $q = 1$ , which means

$$\text{rank } B \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} > \text{boxdim } A_4$$

The rank on the left-hand side is zero, however, and if there exists any period-4 orbit, the filter (3.5) fails this condition. This is consistent with what we have already noticed: in the presence of a period-4 orbit, the map  $F(B, h, g)$  is not one-to-one for any  $h$ .

The filter (3.6) satisfies the above condition as long as there are finitely many period-4 orbits. However, it fails condition A2 for  $p = 4$ ,  $q = 2$ , which requires

$$\text{rank } B \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \\ 1 & 0 \end{pmatrix} > \text{boxdim } A_1$$

This is again consistent with our earlier observation.

Finally, if there exist fixed points, the filter (3.7) fails the condition A1 for  $p = 1$  if there exist fixed points. That is because condition A1 requires

$$\text{rank } B \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} > 2 \cdot \text{boxdim } A_1$$

Since the rank on the left side is zero, the condition fails unless the set of fixed points is empty.

### 3.4. Conditions on Periodic Orbits Which Imply an Immersion

There are also rather obvious situations when certain filters cause  $F(B, h, g)$  to fail to be an immersion. Assume that  $g$  is a diffeomorphism on a circle that has a fixed point  $x$ . Assume that the derivative of  $g$  at  $x$  is  $-2$ . Consider the filter

$$B = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ 0 & \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \end{pmatrix} \quad (3.9)$$

In this case, the map  $F(B, h, g)$  cannot be an immersion at  $x$  for any observation function  $h$ . For a tangent vector  $v$  in  $T_x M = \mathbb{R}^1$ , the derivative map is

$$DF(B, h, g)(x)v = B \begin{pmatrix} \nabla h(x)^T v \\ \nabla h(g(x))^T Dg(x)v \\ \vdots \\ \nabla h(g^{n-1}(x))^T Dg^{n-1}(x)v \end{pmatrix} \\ = \begin{pmatrix} 1 & \vdots & 0 & 0 \\ 0 & \vdots & 1 & 0 \\ 0 & 0 & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \nabla h(x)^T v \\ \nabla h(x)^T (-2v) \\ \nabla h(x)^T (4v) \\ \nabla h(x)^T (-8v) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

so the tangent map of  $F(B, h, g)$  at  $x$  is the zero map.

In the case of an  $m$ -dimensional manifold  $M$  with a fixed point  $x$ , it can be checked that for a filter  $B$  of this type,  $F(B, h, g)$  will fail to be an immersion for all  $h$  as long as the linearization of  $g$  at  $x$  has an eigenvalue of  $-2$ . As in the one-to-one case, the immersion will fail only for periodic points.

To be precise, given numbers  $c_1, \dots, c_r$ , define the  $r \times rp$  matrix

$$D_p^r(c_1, \dots, c_r) = \begin{pmatrix} I_p & \cdots & I_p \\ c_1 I_p & \cdots & c_r I_p \\ \vdots & & \vdots \end{pmatrix} \quad (3.10)$$

where  $I_p$  denotes the  $p \times p$  identity matrix. For a positive integer  $w$ , let  $D_p^w(c_1, \dots, c_r)$  be the matrix formed by the top  $w$  rows of  $D_p^r(c_1, \dots, c_r)$ . If the  $c_i$  are distinct, then  $\text{rank } D_p^w(c_1, \dots, c_r) = \min\{w, rp\}$ .

**Theorem 3.5.** Let  $U$  be an open subset of  $R^n$ , let  $g$  be a smooth diffeomorphism on  $U$ , and let  $A$  be a compact subset of a smooth  $m$ -manifold in  $U$ . Let  $w$  and  $n$  be integers satisfying  $w \geq m \geq 2n$ . Assume that the linearizations  $Dg^p$  of periodic orbits of period  $p$  less than or equal to  $w$  have distinct eigenvalues. Assume that  $B$  is an  $n \times w$  matrix of rank  $n$  which satisfies:

- A3.**  $\text{rank } BD_p^w(\lambda_1, \dots, \lambda_r) > \text{boxdim}(A_p) + r - 1$  for all  $1 \leq p \leq w$ ,  $1 \leq r \leq m$ , and for all subsets  $\lambda_1, \dots, \lambda_r$  of eigenvalues of the linearization at a point in  $A_p$ .

Then for almost every smooth function  $h$ ,  $F(B, h, g)$  is an immersion on  $A$ .

**Remark 3.6.** See Theorem 4.14 for a proof. Note that since  $\text{rank } D_p^w(\lambda_1, \dots, \lambda_r) = \min\{w, rp\}$  for distinct eigenvalues  $\lambda_i$ , it follows that  $\text{rank } BD_p^w = \min\{n, rp\}$  for the original delay coordinate case of  $B = I_n$ , and also for almost every  $n \times w$  matrix  $B$ .

To illustrate, the condition A3 is not satisfied for filter (3.9) when  $g$  has a fixed point with an eigenvalue of  $-2$ , that condition requires that  $\text{rank } BD^*(-2) > 0$ , but

$$BD^*(-2) = B \begin{pmatrix} 1 \\ -2 \\ (-2)^2 \\ (-2)^3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

#### 4. PROOFS

This section contains the proofs of the results stated above. After some fundamental lemmas, we give the proofs of the Whitney forms of the embedding theorems. These follow Lemma 4.1. The proofs of the delay-coordinate forms involving filters, Theorems 3.3 and 3.5, follow immediately from Theorems 4.13 and 4.14, respectively. This section concludes with the proof of Theorems 2.7 and 3.1, which are special cases of Theorems 3.3 and 3.5.

**Lemma 4.1.** Let  $n$  and  $k$  be positive integers,  $y_1, \dots, y_n$  distinct points in  $R^k$ , and  $u_1, \dots, u_n$  in  $R$ ,  $v_1, \dots, v_n$  in  $R^k$ .

1. There exists a polynomial  $h$  in  $k$  variables of degree at most  $n-1$  such that for  $i = 1, \dots, n$ ,  $h(y_i) = u_i$ .
2. There exists a polynomial  $h$  in  $k$  variables of degree at most  $n$  such that for  $i = 1, \dots, n$ ,  $\nabla h(y_i) = v_i$ .

*Proof.* 1. We may assume, by linear change of coordinates, that the first coordinates of  $y_1, \dots, y_n$  are distinct. Then ordinary one-variable interpolation guarantees such a polynomial.

2. First assume  $k = 1$ . There exists a polynomial of degree at most  $n-1$  in one variable that interpolates the data. The antiderivative is the desired polynomial  $h$ .

In the general case, by a linear change of coordinates, we may assume that for each  $j = 1, \dots, k$ , the  $j$ th coordinates of  $y_1, \dots, y_n$  are distinct. The above paragraph shows that for  $j = 1, \dots, k$  there is a polynomial of degree at most  $n$  in the  $j$ th coordinate  $x_j$  whose derivative  $h_{j,x_j}$  interpolates the  $j$ th coordinate of  $u_i$  for  $i = 1, \dots, n$ . The sum of all  $k$  of these polynomials is a polynomial of degree at most  $n$  which satisfies the conclusion.

**Lemma 4.2.** Let  $F(x) = Mx + b$  be a map from  $R^l$  to  $R^n$ , where  $M$  is an  $n \times l$  matrix and  $b \in R^n$ . For a positive integer  $r$ , let  $\sigma_r > 0$  be the  $r$ th largest singular value of  $M$ . Denote by  $B_r$  the ball centered at the origin

of radius  $\rho$  in  $R'$ , and by  $B_\delta$  the ball centered at the origin of radius  $\delta$  in  $R^n$ . Then

$$\frac{\text{Vol}(B_\rho \cap F^{-1}(B_\delta))}{\text{Vol}(B_\rho)} < 2^{n/2} (\delta/\sigma\rho)^r$$

*Proof.* Note that decreasing any singular value of  $M$  does not decrease the left-hand side. Thus we may assume that the singular values of  $M$  satisfy  $\sigma_1 = \dots = \sigma_r = \sigma$ , and  $0 = \sigma_{r+1} = \sigma_{r+2} = \dots$ . Let  $M = V'SU^T$  be the singular value decomposition of  $M$ . Here  $S$  is a diagonal matrix with entries  $s_{11} = \dots = s_{rr} = \sigma$  and all other entries zero.  $V$  is an  $n \times n$  orthogonal matrix, and  $U$  is a  $t \times t$  orthogonal matrix.

Since the columns of  $U$  and  $V$  each form an orthonormal set, we recognize  $MB_\rho = V'SU^TB_\rho$  as an  $r$ -dimensional ball of radius  $\sigma\rho$  lying in  $R^n$ . In fact, the first  $r$  columns of  $V$  magnified by the factor  $\sigma\rho$  are radii which span  $MB_\rho$ .

The set  $F^{-1}(B_\delta) \cap B_\rho$  consists of the vectors in  $B_\rho$  whose image by  $M$  lands in a ball of radius  $\delta$  in  $R^n$ . This is a cylindrical subset of  $B_\rho$  with base dimension  $r$  and base radius  $\delta/\sigma$ . The subset thus has  $t$ -dimensional volume less than  $(\delta/\sigma)^r C_r \rho^{t-r} C_{t-r}$ , where  $C_r = \pi^{r/2}/(r/2)!$  denotes the volume of the  $r$ -dimensional unit ball. The volume of  $B_\rho$  is  $\rho^t C_t$ , so

$$\frac{\text{Vol}(B_\rho \cap F^{-1}(B_\delta))}{\text{Vol}(B_\rho)} < \frac{(\delta/\sigma)^r \rho^{t-r} C_{t-r} C_n}{\rho^t C_t} < 2^{t/2} \left( \frac{\delta}{\sigma\rho} \right)^r$$

**Lemma 4.3.** Let  $S$  be a bounded subset of  $R^d$ ,  $\text{boxdim}(S) = d$ , and let  $G_0, G_1, \dots, G_r$  be Lipschitz maps from  $S$  to  $R^n$ . Assume that for each  $x$  in  $S$ , the  $r$ th largest singular value of the  $n \times r$  matrix

$$M_x = \{G_1(x), \dots, G_r(x)\}$$

is at least  $\sigma > 0$ . For each  $x \in R^d$  define  $G_x = G_0 + \sum_{i=1}^r x_i G_i$ . Then for almost every  $x$  in  $R^d$ , the set  $G_x^{-1}(0)$  has lower box-counting dimension at most  $d - r$ . If  $r > d$ , then  $G_x^{-1}(0)$  is empty for almost every  $x$ .

*Proof.* For a positive number  $\rho$ , define the set  $B_\rho$  to be the ball of radius  $\rho$  centered at the origin in  $R^d$ . For the purposes of proving the theorem, we may replace  $R^d$  by  $B_\rho$ . For the remainder of the proof, we will say that  $G_x$  has some property with probability  $p$  to mean that the Lebesgue measure of the set of  $x \in B_\rho$  for which  $G_x$  has the property is  $p$  times the measure of  $B_\rho$ . For example, if  $x \in S$ , then Lemma 4.2 shows

that  $\|G_z(x)\| = \|G_0(x) + M_z(x)\| \leq \varepsilon$  for  $x \in B_r$  with probability at most  $2^{n/2}(\varepsilon/\sigma\rho)^r$ .

Let  $D > d$  and let  $\varepsilon_0 > 0$  be such that for  $0 < \varepsilon \leq \varepsilon_0$ , the following two facts hold. First,  $S$  can be covered by  $\varepsilon^{-D}$   $k$ -dimensional balls  $B(x, \varepsilon)$  of radius  $\varepsilon$ , centered at  $x \in S$ . Second, by the Lipschitz condition there exists a constant  $C$  such that the image under any  $G_z$ ,  $z \in B_r$ , of any  $\varepsilon$ -ball in  $R^k$  intersecting  $S$  is contained in a  $C\varepsilon$ -ball in  $R^n$ . For the remainder of the proof, we assume  $\varepsilon \leq \varepsilon_0$ .

The probability that the set  $G_z(B(x, \varepsilon))$  contains 0 is at most the probability that  $\|G_z(x)\| < C\varepsilon$ , which is a constant times  $\varepsilon^r$ , since  $\rho$  and  $\sigma$  are fixed. For any positive number  $M$ , the probability that at least  $M$  of the  $\varepsilon^{-D}$  images  $G_z(B(x, \varepsilon))$  contain 0 is at most  $C_1 \varepsilon^{r-D} M$ . Therefore,  $G_z^{-1}(0)$  can be covered by fewer than  $M = \varepsilon^{-b}$  of the  $\varepsilon$ -balls except with probability at most  $C_1 \varepsilon^{b-(D-r)}$ . As long as  $b > D - r$ , this probability can be made as small as desired by decreasing  $\varepsilon$ .

Let  $p > 0$ . There is a sequence  $\{\varepsilon_i\}_{i=1}^\infty$  approaching 0 such that  $G_z^{-1}(0)$  can be covered by fewer than  $\varepsilon_i^{-b}$  balls except for probability at most  $p/2^i$ . Thus, the lower box-counting dimension of  $G_z^{-1}(0)$  is at most  $b$ , except for a probability  $p$  subset of  $z$ . Since  $p > 0$  was arbitrary, lower  $\text{boxdim}(G_z^{-1}(0)) \leq b$  for almost every  $z$ . Finally, since this holds for all  $b > d - r$ , lower  $\text{boxdim}(G_z^{-1}(0)) \leq d - r$ . ■

**Remark 4.4.** In case  $\text{boxdim}(S)$  does not exist, the hypotheses of the lemma can be slightly weakened by allowing  $d$  to be the lower box-counting dimension of  $S$ . A slight adaptation of the proof shows that  $\text{boxdim}$  can be replaced throughout Lemma 4.3 by Hausdorff dimension. In particular, if  $r > \text{HD}(S)$ , then  $G_z^{-1}(0)$  is empty for almost every  $z$  in  $R^l$ .

If in Lemma 4.3 we assume that  $\text{rank}(M_x) \geq d$  for each  $x \in S$  instead of the assumption on the singular values, then  $G_z^{-1}(0)$  is empty for almost every  $z$ . That is because one can apply Lemma 4.3 to the set  $S_n = \{x \in S: r\text{th largest singular value of } M_x > \sigma\}$  to get  $G_z^{-1}(0) \cap S_n = \emptyset$ . Then  $S = \bigcup_{\sigma > 0} S_n$  implies  $G_z^{-1}(0) = \emptyset$ . We state this fact in the next lemma.

**Lemma 4.5.** Let  $S$  be a bounded subset of  $R^k$ ,  $\text{boxdim}(S) = d$ , and let  $G_0, G_1, \dots, G_r$  be Lipschitz maps from  $S$  to  $R^n$ . Assume that for each  $x$  in  $S$ , the rank of the  $n \times r$  matrix

$$M_x = \{G_1(x), \dots, G_r(x)\}$$

is at least  $r$ . For each  $z \in R^r$  define  $G_z = G_0 + \sum_{i=1}^r z_i G_i$ . Then for almost every  $z$  in  $R^r$ , the set  $G_z^{-1}(0)$  is the nested countable union of sets of lower box-counting dimension at most  $d - r$ . If  $r > d$ , then  $G_z^{-1}(0)$  is empty for almost every  $z$ .

**Lemma 4.6.** Let  $A$  be a compact subset of  $R^k$ . Let  $F_0, F_1, \dots, F_t$  be Lipschitz maps from  $A$  to  $R^n$ . For each integer  $r \geq 0$ , let  $S_r$  be the set of pairs  $x \neq y$  in  $A$  for which the  $n \times t$  matrix

$$M_{xy} = \{F_1(x) - F_1(y), \dots, F_t(x) - F_t(y)\}$$

has rank  $r$ , and let  $d_r = \text{lower boxdim}(\overline{S_r})$ . Define  $F_r = F_0 + \sum_{i=1}^r z_i F_i: A \rightarrow R^n$ . Then for  $z = (z_1, \dots, z_r)$  outside a measure zero subset of  $R^r$ , the following hold:

1. If  $d_r < r$  for all integers  $r \geq 0$ , then the map  $F_r$  is one-to-one.
2. If  $d_r \geq r$  for some integer  $r \geq 0$ , then for every  $\delta > 0$ , the lower box-counting dimension of the  $\delta$ -distant self-intersection set  $\Sigma(F_r, \delta)$  is at most  $d_r - r$ .

*Proof.* For  $i = 0, \dots, t$ , define  $G_i(x, y) = F_i(x) - F_i(y)$ . On the set  $S_r$ , the rank of the  $n \times t$  matrix

$$M_{xy} = \{G_1(x, y), \dots, G_t(x, y)\}^t$$

is  $r$ .

If  $r > d_r$ , Lemma 4.5 shows that for almost every  $z \in R^r$ , the origin is not in the image of  $S_r$  under the map  $G_r = G_0 + \sum z_i G_i$ , or equivalently,  $F_r(x) \neq F_r(y)$  for  $x \neq y$  in  $S_r$ . If  $r > d_r$  for all  $r$ , then  $F_r$  is one-to-one, since each pair  $x \neq y$  lies in some  $S_r$ .

If  $r \leq d_r$ , let  $(A \times A)_\delta = \{(x, y) \in A \times A : |x - y| \geq \delta\}$  be the subset of  $\delta$ -distant pairs of points in  $A \times A$ . Since  $(A \times A)_\delta$  is compact for any  $\delta > 0$ , the minimum of the  $n$ th singular value of  $M_{xy}$  in  $(A \times A)_\delta$  is greater than 0. Lemma 4.3 shows that for almost every  $z$ , the origin is in  $G_r(A \times A)_\delta$  for a subset of  $(A \times A)_\delta$  with lower box-counting dimension at most  $d_r - r$ . Therefore the  $\delta$ -distant self-intersection subset  $\Sigma(F_r, \delta)$  of  $A$ , which is the image of this subset under the projection of  $(A \times A)_\delta$  to  $A$ , has dimension at most  $d_r - r$ . ■

**Theorem 4.7.** Let  $A$  be a compact subset of  $R^k$ ,  $\text{lower boxdim}(A) = d$ . If  $n > 2d$ , then almost every linear transformation of  $R^k$  to  $R^n$  is one-to-one on  $A$ .

*Proof.* This follows immediately from Lemma 4.6 and the remark following it. Let  $\{F_i\}$  be a basis for the  $nk$ -dimensional space of linear transformations. For each pair  $x \neq y$ , the vector  $x - y$  can be moved to any direction in  $R^n$  by a linear transformation. In the terminology of Lemma 4.6,  $S_n = A \times A - \Delta$  and  $S_r$  is empty for  $r \neq n$ . Since  $\text{lower boxdim}(\overline{S_n}) = 2d < n$ , almost every  $F_r = \sum z_i F_i$  is one-to-one on  $A$ . ■

**Remark 4.8.** It is interesting that no statement similar to Theorem 4.7 can be made if box-counting dimension is replaced by Hausdorff dimension. In an Appendix to this work provided by I. Kan, examples are constructed of compact subsets  $A$  of any Euclidean space  $R^k$  that have Hausdorff dimension  $d = 0$ , and such that no projection to  $R^n$  for  $n < k$  is one-to-one on  $A$ .

This striking difference between box-counting dimension and Hausdorff dimension is related to the fact that Hausdorff dimension does not work well with products. Extra hypotheses are needed on  $A$ , in particular on the Hausdorff dimension of the product  $A \times A$ , to prove an analogue to Theorem 4.7. For example, Mañé has shown (see ref. 17 and its correction in ref. 9, p. 627) that if  $n > \text{HD}(A \times A) + 1$ , then the conclusion of Theorem 4.7 again holds. Of course, using Lemma 4.3 and Remark 4.4, it turns out that only  $n > \text{HD}(A \times A)$  is required:

**Theorem 4.9.** Let  $A$  be a compact subset of  $R^k$ , and let  $n > \text{HD}(A \times A)$ . Then almost every linear transformation of  $R^k$  to  $R^n$  is one-to-one on  $A$ .

It was shown in ref. 10 that under the hypothesis of Theorem 4.7, almost every *orthogonal* projection is one-to-one (and in fact has a Hölder continuous inverse).

**Definition 4.10.** For a compact differentiable manifold  $M$ , let  $T(M) = \{(x, v) : x \in M, v \in T_x M\}$  be the *tangent bundle* of  $M$ , and let  $S(M) = \{(x, v) \in T(M) : |v| = 1\}$  denote the *unit tangent bundle* of  $M$ .

**Lemma 4.11.** Let  $A$  be a compact subset of a smooth manifold embedded in  $R^k$ . Let  $F_0, F_1, \dots, F_l : R^k \rightarrow R^n$  be a set of smooth maps from an open neighborhood  $U$  of  $A$  to  $R^n$ . For each positive integer  $r$ , let  $S_r$  be the subset of the unit tangent bundle  $S(A)$  such that the  $n \times l$  matrix

$$\{DF_1(x)(v), \dots, DF_l(x)(v)\}$$

has rank  $r$  and let  $d_r = \text{lower boxdim}(\overline{S_r})$ . Define  $F_z = F_0 + \sum_{i=1}^l z_i F_i : U \rightarrow R^n$ . Then the following hold:

1. If  $d_r < r$  for all integers  $r \geq 0$ , then for almost every  $z \in R^l$ , the map  $F_z$  is an immersion on  $A$ .
2. If  $d_r \geq r$  for some  $r \geq 0$ , then for almost every  $z \in R^l$ ,  $F_z$  is an immersion outside a subset of  $A$  of lower boxdim  $\leq d_r - r$ .

*Proof.* For  $i = 0, \dots, l$ , define  $G_i : S(A) \rightarrow R^n$  by  $G_i(x, v) = DF_i(x)v$ . If  $r > d_r$  for all  $r \geq 0$ , then Lemma 4.5 applies to show that for almost every  $z$ ,  $G_z^{-1}(0) \cap S_r$  is the empty set. Since  $S(A)$  is the union of all  $S_r$ ,  $G_z^{-1}(0)$

is empty. Thus, no unit tangent vector is mapped to the origin, and  $F_r$  is an immersion.

In case  $d_r \geq r$  for some  $r$ , there is a positive lower bound on the singular values of the  $G_r$  on  $S(A)$ . Lemma 4.3 implies that there is a subset of unit tangent vectors of lower boxdim  $\leq d_r - r$  that can map to zero. The projection of this subset into  $A$  has lower boxdim  $\leq d_r - r$ . ■

*Proof of Theorems 2.2, 2.3, and 2.10.* Theorem 2.2 is a special case of Theorem 2.3. To prove the latter, we need to show that a prevalent set of maps are one-to-one and immersive.

Let  $F_1, \dots, F_r$  be a basis for the set of linear transformations from  $R^k \rightarrow R^n$ . In the notation of Lemma 4.6, the set  $S_n = A \times A - I$  and  $S_r = \emptyset$  for  $r \neq n$ . Since  $\text{boxdim}(A \times A) = 2d < n$ ,  $F_r$  is one-to-one on  $A$  for almost every  $x \in R^k$ . If any other maps  $F_1, \dots, F_r$  are added, the rank of  $M_x$  cannot drop for any pair  $x \neq y$ , so almost every linear combination of  $F_1, \dots, F_r$  is one-to-one on  $A$ .

The proof of the immersion half uses Lemma 4.11 instead of Lemma 4.6. Since  $\text{boxdim}(A) = d$ ,  $C$  is a subset of a smooth manifold of dimension at most  $d$ , and therefore  $\text{boxdim } S(C) \leq 2d - 1$ . In the notation of Lemma 4.11,  $S_n = S(C)$  and  $S_r = \emptyset$  for  $r \neq n$ . Since  $n > 2d > 2d - 1 = \text{boxdim } S_n$ , the proof follows from Lemma 4.11.

The proof of Theorem 2.10 is similar, except that the second part of the conclusions of Lemmas 4.6 and 4.11 are used. For example, in the use of Lemma 4.6,  $S_n = A \times A - I$  and  $S_r = \emptyset$  for  $r \neq n$  as before, but now  $\text{boxdim}(A \times A) = 2d \geq n$ . Thus for each  $\delta > 0$ , for almost every  $F_r$ , the  $\delta$ -distant self-intersection set  $\Sigma(F_r, \delta)$  has lower box-counting dimension at most  $2d - n$ . The immersion half is again analogous. ■

**Definition 4.12.** Let  $U$  be an open subset of  $R^k$ , let  $g: U \rightarrow U$  be a map, and let  $h: U \rightarrow R$  be a function. Let  $w^- < w^+$  be integers and set  $w = w^+ - w^- + 1$ . For  $1 \leq i \leq w$ , set  $g_i = g^{w^- + i - 1}$ , so that  $g_1 = g^{w^-}$  and  $g_w = g^{w^+}$ . Let  $B$  be an  $n \times w$  matrix. Define the *filtered delay-coordinate map*

$$F_w^{w^+}(B, h, g): U \rightarrow R^n$$

by

$$\begin{aligned} F_w^{w^+}(B, h, g)(x) &= B(h(g_1(x)), h(g_2(x)), \dots, h(g_w(x)))^T \\ &= B(h(g^{w^-}(x)), \dots, h(g^{w^+}(x)))^T \end{aligned}$$

Theorems 2.7, 3.1, 3.3, and 3.5 are corollaries of the next two results, for which we will use the following notation. Let  $g$  denote a smooth diffeomorphism on an open neighborhood  $U$  in  $R^k$ . Let  $h_1, \dots, h_r$  be a basis

for the polynomials in  $k$  variables of degree at most  $2w$ . For a smooth function  $h_0$  on  $R^k$  and for  $x \in R^t$ , define  $h_x = h_0 + \sum_{i=1}^t x_i h_i$ . For each positive integer  $p$ , denote by  $A_p$  the set of period- $p$  points of  $g$  lying on  $A$ . That is,  $A_p = \{x \in A: g^p(x) = x\}$ . Let the matrices  $C_{pq}^w$  be as in Theorem 3.3.

**Theorem 4.13.** Let  $g$  be a smooth diffeomorphism on an open neighborhood  $U$  of  $R^k$ , and let  $A$  be a compact subset of  $U$ ,  $\text{boxdim}(A) = d$ . Let  $n$  and  $w^+ < w^-$  be integers,  $n \leq w = w^+ - w^- + 1$ . Assume that the  $n \times w$  matrix  $B$  satisfies:

**A1.**  $\text{rank } BC_{p0}^w > 2 \cdot \text{boxdim}(A_p)$  for all  $1 \leq p \leq w$ .

**A2.**  $\text{rank } BC_{pq}^w > \text{boxdim}(A_p)$  for all  $1 \leq q < p \leq w$ .

Let  $h_0, \dots, h_t$  be a basis for the polynomials in  $k$  variables of degree at most  $2w$ . Then for any smooth function  $h_0$  on  $R^k$ , and for almost every  $x \in R^t$ , the following hold:

1. If  $n > 2d$ , then  $F(B, h_x, g): U \rightarrow R^n$  is one-to-one on  $A$ .
2. If  $n \leq 2d$ , then for every  $\delta > 0$ , the  $\delta$ -distant self-intersection set  $\Sigma(F(B, h_x, g), \delta)$  has lower box-counting dimension at most  $2d - n$ .

*Proof.* For  $i = 1, \dots, t$  define

$$F_i(x) = B \begin{pmatrix} h_i(g_1(x)) \\ \vdots \\ h_i(g_n(x)) \end{pmatrix}$$

By definition,  $F(B, h_x, g) = \sum_{i=1}^t F_i$ . To use Lemma 4.6, we need to check for each  $x \neq y$  the rank of the matrix

$$M_{x,y} = (F_1(x) - F_1(y), \dots, F_t(x) - F_t(y))$$

which can be written as

$$B \begin{pmatrix} h_1(g_1(x)) - h_1(g_1(y)) & \cdots & h_t(g_1(x)) - h_t(g_1(y)) \\ \vdots & & \vdots \\ h_1(g_n(x)) - h_1(g_n(y)) & \cdots & h_t(g_n(x)) - h_t(g_n(y)) \end{pmatrix} = BJH$$

where

$$H = \begin{pmatrix} h_1(z_1) & \cdots & h_t(z_1) \\ \vdots & & \vdots \\ h_1(z_q) & \cdots & h_t(z_q) \end{pmatrix}$$

$q \leq 2w$ , the  $z_i$  are distinct, and  $J = J_{w,q}$  is a  $w \times q$  matrix each of whose rows consists of zeros except for one 1 and one  $-1$ . By part 1 of Lemma 4.1, the

rank of  $H$  is  $q$ . We divide the study of the rank of  $M_{vv} = BJH$  into three cases.

**Case 1:**  $x$  and  $y$  are not both periodic with period  $\leq w$ .

In this case,  $J_{vv}$  is upper or lower triangular, and  $\text{rank}(J_{vv}) = w$ . Since  $B, J$ , and  $H$  are onto linear transformations, the product  $BJH$  is onto and has rank  $n$ . The set of pairs  $x \neq y$  of case 1 has box-counting dimension at most  $2d$ , and  $\text{rank}(M_{vv}) = n$ . If  $g$  has no periodic points of period  $\leq w$ , we are done, and conclusion 1 (respectively, 2) of Lemma 4.6 implies conclusion 1 (resp., 2) of the theorem.

The remaining two cases are necessary to deal with periodic points of period  $\leq w$ . We show that conclusion 1 of Lemma 4.6 applies in both cases.

**Case 2:**  $x$  and  $y$  lie in distinct periodic orbits of period  $\leq w$ .

Assume  $p$  and  $q$  are minimal such that  $g^p(x) = x$ ,  $g^q(y) = y$ , and that  $1 \leq q \leq p \leq w$ . In this case the matrix  $J_{vv}$  contains a copy of  $C_{pq}^w$ . Since  $H$  is onto,  $\text{rank } M_{vv} = \text{rank } BJ_{vv}H = \text{rank } BJ_{vv}$ . By hypothesis,  $\text{rank } BJ_{vv} > \text{rank } BC_{pq}^w > 2 \cdot \text{boxdim } A_p$ , which is the box-counting dimension of the set of pairs treated in case 2. By Lemma 4.6, for almost every  $x \in R'$ ,  $F_p(x) \neq F_q(y)$  for every such pair  $x \neq y$ .

**Case 3:** Both  $x$  and  $y$  lie in the same periodic orbit of period  $\leq w$ .

Assume  $p$  and  $q$  are minimal such that  $g^p(x) = x$ ,  $g^q(x) = y$ , and that  $1 \leq q < p \leq w$ . Since  $x$  and  $y$  lie in the same periodic orbit, the column space of  $J_{vv}$  contains the column space of  $C_{pq}^w$ . Thus,  $\text{rank } BJ_{vv}H = \text{rank } BJ_{vv} > \text{rank } BC_{pq}^w > \text{boxdim } A_p$ , which is the dimension of the pairs  $x \neq y$  of case 3. Now Lemma 4.6 applies to give the conclusion. ■

**Theorem 4.14.** Let  $g$  be a smooth diffeomorphism on an open neighborhood  $U$  in  $R^k$ , and let  $A$  be a compact subset of a smooth  $m$ -manifold in  $U$ . Assume that the linearizations of periodic orbits of period less than  $w$  have distinct eigenvalues. Let  $n \leq w$  be positive integers as in Theorem 4.13, and assume that the  $n \times w$  matrix  $B$  satisfies:

- A3.**  $\text{rank } BD_p^w(\lambda_1, \dots, \lambda_r) > \text{boxdim}(A_p \div r - 1)$  for all  $1 \leq p < w$ ,  $1 \leq r \leq m$ , and for all subsets  $\lambda_1, \dots, \lambda_r$  of eigenvalues of the linearization  $Dg^p$  at a point in  $A_p$ .

Let  $h_1, \dots, h_l$  be a basis for the polynomials in  $k$  variables of degree at most  $2w$ . Then for any smooth function  $h_0$  on  $R^k$ , and for almost every  $x \in R'$ , the following hold:

1. If  $n \geq 2m$ , then  $F(B, h_x, g): U \rightarrow R^n$  is an immersion on  $A$ .
2. If  $n < 2m$ , then  $F(B, h_x, g)$  is an immersion outside an exceptional subset of  $A$  of dimension at most  $2m - n - 1$ .

*Proof.* To apply Lemma 4.11, we need to check the rank of the  $n \times r$  matrix

$$(DF_1(x)(v), \dots, DF_r(x)(v)) \quad (4.1)$$

for each  $(x, v)$  in the unit tangent bundle  $S(A)$ . For a given observation function  $h$ , the derivative of  $F(B, h, g)$  is

$$DF(B, h, g)(x)v = B \begin{pmatrix} \nabla h(g^{n-1}(x))^T Dg^{n-1}(x)v \\ \vdots \\ \nabla h(g^{n-r}(x))^T Dg^{n-r}(x)v \end{pmatrix}$$

If  $x$  is not a periodic point of period less than  $w$ , then  $g^{n-1}(x), \dots, g^{n-r}(x)$  are distinct points. The facts that  $g$  is a diffeomorphism and  $v \neq 0$  imply that  $Dg^i(x)v \neq 0$  for all  $i$ . Therefore by Lemma 4.1, part 2, the set of vectors  $\{DF(B, h, g)(x)v : v \in R^r\}$  spans  $R^n$ . In the notation of Lemma 4.11, the subset  $S_n$  contains all points of  $S(A)$  that are not periodic with period less than  $w$ , and  $d_n = \text{lower boxdim}(\overline{S_n}) \leq 2m - 1$ . If  $g$  has no periodic points of period less than  $w$ , the proof is finished, by Lemma 4.11.

If  $x$  is a periodic point of period  $p < w$ , then

$$DF(B, h, g)(x)v = B \begin{pmatrix} H_1^T w_1 \\ \vdots \\ H_p^T w_p \\ H_1^T D_1 w_1 \\ \vdots \\ H_p^T D_p w_p \\ H_1^T D_1^2 w_1 \\ \vdots \\ H_1^T D_1^{p-1} w_1 \end{pmatrix}$$

where

$$x_i = g^{n-i}(x) = x_{p+i}$$

$$H_i = \nabla h(x_i)$$

$$w_i = Dg(x_{i-1}) \cdots Dg(x_1) Dg^{n-1}(x)v$$

$$D_i = Dg(x_{i-1}) \cdots Dg(x_1) Dg(x_p) \cdots Dg(x_1)$$

Each matrix  $D_i$  has the same set of eigenvalues  $\lambda_1, \dots, \lambda_m$ , and by hypothesis, they are distinct. If  $u_1, \dots, u_m$  is a spanning set of eigenvectors for  $D_1$ , then it checks that  $u_{ij} = Dg(x_{i-1}) \cdots Dg(x_1) u_j$  for  $1 \leq i \leq p$ ,  $1 \leq j \leq m$  defines a spanning set  $\{u_{i1}, \dots, u_{im}\}$  of eigenvectors for  $D_i$ . Thus, if

$w_i = \sum_{j=1}^m a_j u_{ij}$  is the eigenvector expansion of  $w_i$ , then the eigenvector expansion of  $w_i$  is  $\sum_{j=1}^m a_j u_{ij}$ , which has the same coefficients.

Thus  $DF(B, h, g)(x)v$  can be written as  $B$  times the  $w$ -vector

$$\begin{pmatrix} 1 & \dots & 1 \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \\ 0 & \dots & 0 \\ \lambda_1 & \dots & \lambda_m \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \\ 0 & \dots & 0 \\ \lambda_1^2 & \dots & \lambda_m^2 \\ \vdots & & \vdots \end{pmatrix} \begin{pmatrix} a_1 u_{11}^T \\ \vdots \\ a_m u_{m1}^T \end{pmatrix} H_1 + \dots + \begin{pmatrix} 0 & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \\ 1 & \dots & 1 \\ 0 & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \\ \lambda_1 & \dots & \lambda_m \\ 0 & \dots & 0 \\ \vdots & & \vdots \end{pmatrix} \begin{pmatrix} a_1 u_{p1}^T \\ \vdots \\ a_m u_{pm}^T \end{pmatrix} H_p \quad (4.2)$$

To find the rank of the matrix (4.1) for  $(x, v)$  where  $x$  is periodic, we need to find the span of  $B$  times the vectors (4.2) for  $h = h_x = \sum x_i h_i$ ,  $x \in R^l$ . Assume that the eigenvector expansion of  $v$  has exactly  $r$  nonzero coefficients  $a_1, \dots, a_r$ . By Lemma 4.1, part 2, the set of vectors  $\{\nabla h_x(x_i) : x \in R^l\}$  spans  $R^k$ . Then because the  $u_j$ ,  $1 \leq j \leq m$ , are linearly independent, the vectors of form (4.2) span a space of dimension  $\min\{w, rp\}$  as  $x$  spans  $R^l$ .

Therefore, for this  $(x, v)$ , the span of the vectors (4.1) has dimension equal to the rank of  $BD_p^w(\lambda_1, \dots, \lambda_m)$ . By hypothesis, the boxdim of such pairs  $(x, v)$  in  $S(\mathcal{A})$  is  $\text{boxdim}(\mathcal{A}_p) + r - 1$ . By hypothesis, the rank of the  $n \times r$  matrix (4.1) is strictly larger, so that Lemma 4.11 applies to give the conclusion.

**Proof of Theorem 2.7.** Apply Theorems 3.3 and 3.5 with  $B = I_n$ . According to Remarks 3.4 and 3.6, the conditions A1-A3 translate to  $p > 2 \cdot \text{boxdim}(\mathcal{A}_p)$ ,  $p/2 > \text{boxdim}(\mathcal{A}_p)$ , and  $\min\{n, rp\} > \text{boxdim}(\mathcal{A}_p) + r - 1$ , respectively, for  $1 \leq p \leq n$  and  $1 \leq r \leq m$ . Thus, the hypothesis  $\text{boxdim}(\mathcal{A}_p) < p/2$  guarantees that A1-A3 hold.

**Proof of Theorem 3.1.** Since  $\mathcal{A}_p$  is empty for  $1 \leq p \leq w$ , the conditions A1-A3 of Theorems 3.3 and 3.5 are satisfied vacuously.

## APPENDIX. HAUSDORFF DIMENSION-ZERO SETS WITH NO ONE-TO-ONE PROJECTIONS

Ittai Kan<sup>4</sup>

The purpose of this Appendix is to construct a Cantor set  $C \subset R^m$  whose Hausdorff dimension is zero and which has the property that every projection of rank less than  $m$  is not one-to-one when restricted to  $C$ .

**Definition A.1.** The Hausdorff  $s$ -dimensional outer measure of a set  $K$  is

$$\mathcal{H}^s(K) = \lim_{\delta \rightarrow 0} \inf \sum_{i=1}^r |U_i|^s$$

where the infimum is taken over all covers  $\{U_i\}$  of  $K$  with the diameters of the  $U_i$  uniformly less than  $\delta$ . The Hausdorff dimension of a nonempty set  $K$  is the unique value of  $s$  such that

$$\mathcal{H}^t(K) = \infty \text{ if } t < s \quad \text{and} \quad \mathcal{H}^t(K) = 0 \text{ if } t > s$$

**Example A.2.** We construct the subset  $C$  of  $R^m$  as the union of two sets  $A = \bigcup_{n=1}^m A_n$  and  $B = \bigcup_{n=1}^m B_n$  each of Hausdorff dimension zero, with the property that for any projection  $P$  of rank less than  $m$  the images under  $P$  of  $A$  and  $B$  intersect, and thus  $P$  is not injective when restricted to  $C$ .

The set  $A_n$  lies on a face of the unit  $m$ -cube and  $a = (a_1, a_2, \dots, a_m)$  is in  $A_n$  if it satisfies the following restrictions on the binary expansion  $a_i = a_i^1 a_i^2 a_i^3 \dots$  of its coordinates:

1. If  $i = n$ , then  $a_i = 0$ .
2. If  $i \neq n$  and  $k \geq 0$ , then either (a)  $a_i^l = 0$  for all  $l \in (M_{2k}, M_{2k+1}]$ ; or (b)  $a_i^l = 1$  for all  $l \in (M_{2k}, M_{2k+1}]$ .

Here the sequence  $0 = M_0 < M_1 < M_2 \dots$  increases sufficiently rapidly so that  $\lim_{j \rightarrow \infty} (M_{j+1}, M_j) = \infty$ . If  $i \neq n$ , then the orthogonal projection of  $A_n$  on the  $i$ th coordinate axis is a Cantor set which can be covered by  $2^{r_k}$  intervals of length  $2^{-M_{2k+1}}$ , where  $r_k = k + \sum_{j=1}^k (M_{2j} - M_{2j-1})$ . Thus,  $A_n$  can be covered by  $2^{(m-1)r_k}$  cubes with edges of length  $2^{-M_{2k+1}}$ . Since  $r_k \leq M_{2k}$ , we see that  $\lim_{k \rightarrow \infty} (m-1)r_k M_{2k+1}^{-1} = 0$  and both the lower box-counting and Hausdorff dimensions of  $A_n$  are zero. Since  $A$  is the union of  $m$  copies of  $A_n$ , we see that both the lower box-counting and Hausdorff dimensions of  $A$  are zero.

<sup>4</sup> Department of Mathematical Sciences, George Mason University, Fairfax, Virginia 22030.

The set  $B_n$  lies on a face of the unit  $m$ -cube opposite  $A_n$  and  $b$  is in  $B_n$  if it satisfies the following restrictions on the binary expansion of its coordinates:

1. If  $k=n$ , then  $b'_i = 1$ .
2. If  $i \neq n$  and  $k \geq 0$ , then either (a)  $b'_i = 0$  for all  $i \in (M_{2k+1}, M_{2k+2}]$ ; or (b)  $b'_i = 1$  for all  $i \in (M_{2k+1}, M_{2k+2}]$ .

Here  $\{M_i\}$  is as above. The lower box-counting and Hausdorff dimensions of  $B$  are zero. The Hausdorff dimension of  $C = A \cup B$  is zero.

Let  $P$  denote a projection of rank less than  $m$ . Let  $v = (v_1, v_2, \dots, v_m)$  in the null space of  $P$  be chosen so that  $|v_i| \leq 1$  for all  $i$  and  $v_n = 1$  for some particular  $n$ . We now show that  $P$  restricted to  $C$  is not injective by finding some  $b \in B_n$  and  $a \in A_n$  such that  $v = b - a$ . Using the binary expansion coordinate notation, we define  $a$  and  $b$  as follows:

1. If  $i = n$ , then  $a'_i = 0$  and  $b'_i = 1$ .
2. If  $i \neq n$  and  $k \geq 0$ , then (a)  $a'_i = 0$  and  $b'_i = v'_i$  for all  $i \in (M_{2k}, M_{2k+1}]$ ; and (b)  $a'_i = (v'_i + 1) \bmod 2$  and  $b'_i = 1$  for all  $i \in (M_{2k+1}, M_{2k+2}]$ .

Clearly we have  $v = b - a$  and by the definition of  $A_n$  and  $B_n$  we also have  $a \in A_n$  and  $b \in B_n$ . ■

## ACKNOWLEDGMENTS

The research of T.S. and J.A.Y. was supported by the Applied and Computational Mathematics Program of DARPA, that of J.A.Y. additionally by AFOSR and the U. S. Department of Energy (Basic Energy Sciences), and that of M.C. by grants to the Santa Fe Institute, including core funding from the John D. and Catherine T. MacArthur Foundation, the National Science Foundation, and the U. S. Department of Energy.

## REFERENCES

1. H. Abarbanel, R. Brown, and J. Kadtko, Prediction in chaotic nonlinear systems: Methods for time series with broadband Fourier spectra, preprint.
2. A. M. Albano, J. Muench, C. Schwartz, A. Mees, and P. Rapp, Singular value decomposition and the Grassberger-Procaccia algorithm, *Phys. Rev. A* 38:3017-3026 (1988).
3. V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations* (Springer-Verlag, New York, 1983).
4. R. Badii, G. Broggi, B. Derighetti, M. Ravani, S. Ciliberto, A. Politi, and M. A. Rubio, Dimension increase in filtered chaotic signals, *Phys. Rev. Lett.* 60:979-982 (1988).
5. D. S. Broomhead and G. P. King, Extracting qualitative dynamics from experimental data, *Physica* 20D:217-236 (1986).

6. M. Casdagli, Nonlinear prediction of chaotic time series, *Physica* 35D:335-356 (1989).
7. M. Casdagli, S. Eubank, D. Farmer, and J. Gibson, State-space reconstruction in the presence of noise, preprint.
8. W. Ditto, S. Raueo, and M. Spano, Experimental control of chaos, *Phys. Rev. Lett.* 65:3211-3214 (1990).
9. J.-P. Eckmann and D. Ruelle, Ergodic theory of chaos and strange attractors, *Rev. Mod. Phys.* 57:617-656 (1985).
10. A. Eden, C. Foias, B. Nicolaenko, and R. Temam, Hölder continuity for the inverse of Mañé's projection, *Comptes Rendus*, to appear.
11. K. Falconer, *Fractal Geometry* (Wiley, New York, 1990).
12. J. D. Farmer and J. Sidorowich, Predicting chaotic time series, *Phys. Rev. Lett.* 59:845-848 (1987).
13. J. D. Farmer and J. Sidorowich, Exploiting chaos to predict the future and reduce noise, Technical Report LA-UR-88-901, Los Alamos National Laboratory (1988).
14. G. Golub and C. Van Loan, *Matrix Computations*, 2nd ed. (Johns Hopkins University Press, Baltimore, Maryland, 1989).
15. E. Kostelich and J. Yorke, Noise reduction: Finding the simplest dynamical system consistent with the data, *Physica* 41D:183-196 (1990).
16. E. Kostelich and J. Yorke, Noise reduction in dynamical systems, *Phys. Rev.* 38:1649-1652 (1988).
17. R. Mañé, On the dimension of the compact invariant sets of certain nonlinear maps, in *Lecture Notes in Mathematics*, No. 898 (Springer-Verlag, 1981).
18. P. Marteau and H. Abarbanel, Noise reduction in chaotic time series using scaled probabilistic methods, preprint.
19. P. Mattila, Hausdorff dimension, orthogonal projections and intersections with planes, *Ann. Acad. Sci. Fenn. Math.* 1:227-224 (1975).
20. F. Mitschke, M. Möller, and W. Lange, Measuring filtered chaotic signals, *Phys. Rev.* 37:4518-4521 (1988).
21. N. Packard, J. Crutchfield, D. Farmer, and R. Shaw, Geometry from a time series, *Phys. Rev. Lett.* 45:712 (1980).
22. W. Rudin, *Real and Complex Analysis*, 2nd ed. (McGraw-Hill, New York, 1974).
23. J.-C. Roux and H. Swinney, Topology of chaos in a chemical reaction, in *Nonlinear Phenomena in Chemical Dynamics*, C. Vidal and A. Pacault, eds. (Springer, Berlin, 1981).
24. B. Hunt, T. Sauer and J. Yorke, Prevalence: A translation-invariant "almost every" on infinite-dimensional spaces, preprint.
25. T. Sauer and J. Yorke, Statistically self-similar sets, preprint.
26. J. Sommerer, W. Ditto, C. Grebogi, E. Ott, and M. Spano, Experimental confirmation of the theory for critical exponents of crises, *Phys. Lett. A* 153:105-109 (1991).
27. F. Takens, Detecting strange attractors in turbulence, in *Lecture Notes in Mathematics*, No. 898 (Springer-Verlag, 1981).
28. J. Townshend, Nonlinear prediction of speech signals, preprint.
29. H. Whitney, Differentiable manifolds, *Ann. Math.* 37:645-680 (1936).
30. J. Yorke, Periods of periodic solutions and the Lipschitz constant, *Proc. Am. Math. Soc.* 22:509-512 (1969).

## A numerical procedure for finding accessible trajectories on basin boundaries\*

Helena E Nusse†§ and James A Yorke†‡

†Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA

‡Department of Mathematics, University of Maryland, College Park, MD 20742, USA

Received 29 January 1991

Accepted by J Sinai

**Abstract.** In dynamical systems examples are common in which two or more attractors coexist, and in such cases the basin boundary is non-empty. The basin boundary is either smooth or fractal (that is, it has a Cantor-like structure). When there are horseshoes in the basin boundary, the basin boundary is fractal. A relatively small subset of a fractal basin boundary is said to be 'accessible' from a basin. However, these accessible points play an important role in the dynamics and, especially, in showing how the dynamics change as parameters are varied. The purpose of this paper is to present a numerical procedure that enables us to produce trajectories lying in this accessible set on the basin boundary, and we prove that this procedure is valid in certain hyperbolic systems.

AMS classification scheme numbers: 58F12, 58F13, 65Q05

### 1. Introduction

Dynamical systems often have quite different behaviour in different open sets, each open set having its own attractor. These open sets may be the basins of attractors. We are interested in the boundary on the common boundary between such open sets. The common behaviour may be either smooth or fractal. A point  $p$  on the boundary of an open set  $U$  is *accessible* from  $U$  if there is a curve lying in  $U \cup \{p\}$  which ends on  $p$ . The basin boundary is the set of all points on the boundary of a basin of attraction such that each open neighbourhood of  $p$  intersects at least two different basins of attraction [GOY1]. If the basin boundary is smooth, then each point on the basin boundary is accessible from two basins. In particular, if the basin boundary is a curve, then all of its points are accessible. When the basin boundary is

\* Research in part supported by AFOSR, and by DARPA under the Applied & Computational Mathematics Program.

§ Permanent address: Rijksuniversiteit Groningen, Fac. Economische Wetenschappen, WSN-gebouw, Postbus 800, NL-9700 AV Groningen, The Netherlands.

fractal, only a relatively small subset of the basin boundary consists of accessible points, and generally no points that are accessible from a basin will be accessible from another basin. A collection of papers have assumed that investigators can produce accessible trajectories on basin boundaries [AS], [AY], [GOY1], [HJ], but no rigorous procedures have been presented. For more details, see the discussion in section 6.

Studying dynamical systems, one often observes transient chaotic behaviour, apparently due to the presence of horseshoes. It is well known [MGOY] that transient chaos is present whenever there is a fractal basin boundary separating the basins of two or more attractors. For example, for suitably chosen parameter values, the Hénon map has attracting periodic orbits with period 3 and 5, and also a non-attracting chaotic invariant set in the basin boundary, and one observes that the duration of the transient chaotic behaviour of many trajectories is rather short before they settle down to one of these two periodic attractors. Other famous examples with chaotic transients, due to a bounded non-attracting invariant chaotic set in the basin boundary, are the forced damped pendulum and the forced Duffing equation. Transient chaos is also present if there is a chaotic invariant set in the interior of the closure of the basin. In this case, the basin boundary can be either fractal or smooth [KG], [NY1], [NY2].

Let  $M$  be a smooth  $d$ -dimensional manifold without boundary with  $d \geq 2$ , and let  $F$  be a  $C^3$ -diffeomorphism from  $M$  to itself. For  $x, y$  in  $M$  we denote by  $\rho(x, y)$  the distance between  $x$  and  $y$ . A set  $S \subset M$  is *positively invariant* if  $F(S) \subset S$ , and is *invariant* if  $F(S) = S$ . For  $x \in M$  and a closed set  $S \subset M$ , we write  $\rho(x, S) = \min\{\rho(x, y) : y \in S\}$ . An *attractor*  $A$  is an invariant compact set in  $M$  such that: (1) there exists an open neighbourhood  $U$  of  $A$  such that for each  $x \in U$  the distance  $\rho(F^n(x), A) \rightarrow 0$  when  $n \rightarrow \infty$ ; and (2) there is a point  $x \in A$  such that the closure of the trajectory  $\{F^n(x)\}_{n \geq 0}$  equals  $A$ . A *generalized attractor* is the union of finitely many attractors. We say a *region* is an open and bounded set in  $M$ ; a *transient region* is a region that contains no attractor. For an attractor (or a generalized attractor)  $A$  we say, the *domain of attraction* of  $A$  is the set of all points  $x$  in  $M$  for which  $\rho(F^n(x), A) \rightarrow 0$  as  $n \rightarrow \infty$ . The *basin boundary* is the set of all points  $x \in M$  for which each open neighbourhood has a non-empty intersection with at least two different domains of attraction, see [GOY1]. In the literature, for an attractor  $A$  the notions 'domain of attraction of  $A$ ' and 'basin of  $A$ ' are often equivalent. On the other hand, in other studies of dynamical systems, the notion 'basin of  $A$ ' is defined as the region in  $M$  that is the interior of the closure of the domain of attraction of  $A$ . Therefore, for an attractor (or generalized attractor)  $A$  we define  $\text{basin}\{A\}$  to be the interior of the closure of the domain of attraction of  $A$ . We would like to emphasize that  $\text{basin}\{A\}$  is associated with attractor  $A$  and may include Cantor sets of curves that are not in the domain of attraction of  $A$ ; that is, the trajectories of all the points on these curves will not converge to the attractor  $A$ . In the forced pendulum example in section 3 we show numerically that  $\text{basin}\{A\}$  does include such an invariant Cantor set of curves.

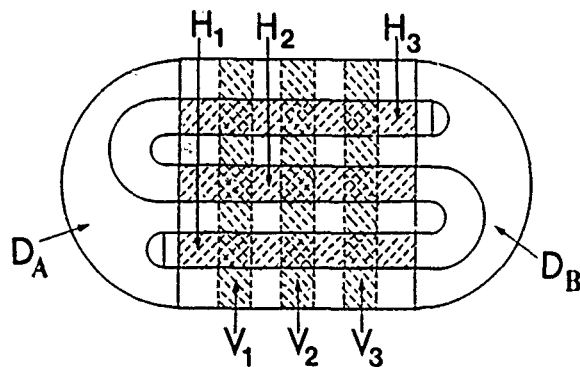
We will be studying transient regions in cases where the trajectory through almost every initial point eventually leaves the region. We investigate special trajectories that remain in such a transient region for all positive time. In [BGOYY], [GNOY] a numerical method (involving the bisection procedure) for finding trajectories on the basin boundary was presented. The papers [NY1], [NY2] introduced the PIM triple (refinement) procedure and the accessible PIM triple

(refinement) procedure. Both these refinement procedures enable us to obtain numerical trajectories and accessible numerical trajectories respectively, that stay (for positive time) in a specified transient region in  $M$ . In [NY2] these two refinement procedures were shown to be valid for uniformly saddle-hyperbolic dynamical systems, for which the dimension of the unstable manifold of any nonwandering point in the transient region was assumed to be one dimensional.

Let  $R$  be a transient region for  $F$ . The *stable set*  $S(R)$  of  $F$  is  $\{x \in R : F^n(x) \in R \text{ for } n=0, 1, 2, \dots\}$ ; the *unstable set*  $U(R)$  of  $F$  is  $\{x \in R : F^{-n}(x) \in R \text{ for } n=0, 1, 2, \dots\}$ . The set of points  $x$  for which  $F^n(x)$  is in  $R$  for all integers  $n$  is called the *invariant set*  $\text{Inv}(R)$  of  $F$  in  $R$ , that is,  $\text{Inv}(R) = S(R) \cap U(R)$ . A component of  $S(R)$  (resp.  $U(R)$ ), which contains a point of  $\text{Inv}(R)$  is called a *stable* (resp. *unstable*) *segment*. We call  $\text{Inv}(R)$  a *chaotic saddle* when it includes a Cantor set. These notions are illustrated in the following example.

**Example 1.** An S-shaped horseshoe map is an invertible map that squeezes, stretches and folds a rectangle into an S-shape area as illustrated in the figure below. We consider the S-shaped horseshoe map  $g$ , which is defined on a neighbourhood of a compact, connected set  $W$ , where  $W$  is the union of a rectangle  $E$  and the two half disks  $D_A$  and  $D_B$  as indicated in the figure. Assume (1)  $g$  maps  $W$  into its interior, (2) the intersection  $g(E) \cap E$  consists of three horizontal strips, say  $H_1$ ,  $H_2$  and  $H_3$ , and (3) the half disks  $D_A$  and  $D_B$  include fixed point attractors  $A$  and  $B$  respectively. Let  $V_1$ ,  $V_2$  and  $V_3$  be the vertical strips in  $E$  (stretching the full width of  $E$ ) such that  $V_i = H_i$ ,  $1 \leq i \leq 3$ ; see figure 1.

It is well known, see e.g. Guckenheimer and Holmes [GH], that under reasonable assumptions, almost every point will be attracted to either  $A$  or  $B$ , the stable set  $S(E)$  of  $g$  with respect to  $E$  is a Cantor set of vertical curves, and the unstable set  $U(E)$  of  $g$  with respect to  $E$  is a Cantor set of horizontal curves. All components of  $S(E)$  are stable segments, and all components of  $U(E)$  are unstable segments. The intersection  $C$  of the stable set  $S(E)$  with the unstable set  $U(E)$  in  $E$  is a chaotic saddle. Note that all the points on the chaotic saddle  $C$  stay in the box  $E$  for all time under all forward and all backward iterates of the map  $g$ . The set of points in  $E$  that are on the basin boundary is the stable set  $S(E)$ , and the basin boundary of  $g$  is fractal. One might choose the transient region  $R$  to be the interior of  $W$  minus two small closed balls that are centred at the attractors  $A$  and  $B$ .

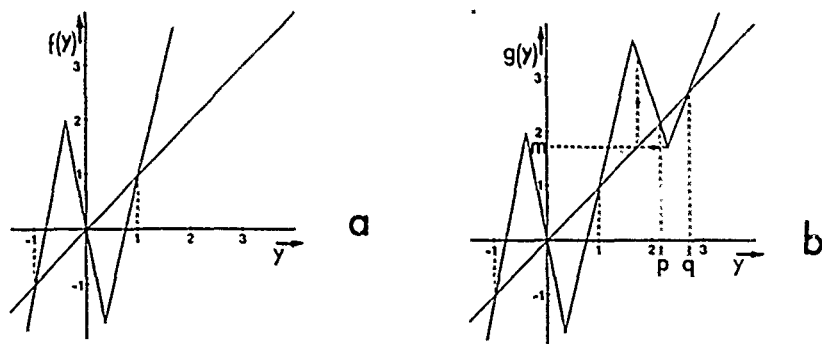


**Figure 1.** S-shape horseshoe map: vertical strips in the rectangle  $E$  are mapped into horizontal strips in  $E$ , namely  $F(V_1) = H_1$ ,  $F(V_2) = H_2$ , and  $F(V_3) = H_3$ . The half disks  $D_A$  and  $D_B$  each contains a fixed point attractor, and each is mapped into its interior.

We assume throughout that (1) for the transient region  $R$  the set  $\text{Inv}(R)$  is non-empty, and (2) there exist two generalized attractors  $A$  and  $B$ , and each point in  $R$  that escapes from  $R$  under iteration of the map  $F$  is either in  $\text{basin}(A)$  or in  $\text{basin}(B)$ , and the basin boundary is the common boundary of  $\text{basin}(A)$  and  $\text{basin}(B)$ .

We will refer to  $R \setminus S(R)$ , the complement of the stable set  $S(R)$  in the transient region  $R$ , as the *transient set*. Recall that a point  $p$  in  $S(R)$  is *accessible* from an open set  $V$  if there is a continuous curve  $K$  ending at  $p$  such that  $K \setminus \{p\}$  is in  $V$ . We investigate the cases where  $V$  is either  $\text{basin}(A)$  (or  $\text{basin}(B)$ ) or is the transient set  $R \setminus S(R)$ . In this paper we emphasize points accessible from  $\text{basin}(A)$  rather than from  $\text{basin}(B)$ , just to simplify notation. Obviously, if a point  $p$  in  $S(R)$  is accessible from the transient set  $R \setminus S(R)$  and  $p$  is on the basin boundary, then  $p$  is accessible from either  $\text{basin}(A)$  or  $\text{basin}(B)$ . On the other hand,  $S(R)$  can contain points which are not in the basin boundary, and such points might be so numerous that they block the access to the basin boundary, that is, every curve in  $\text{basin}(A)$  that goes to an accessible point would necessarily pass through points of  $S(R)$ . Thus no points of the basin boundary would be accessible from  $R \setminus S(R)$ . Naturally  $S(R)$  would have its own accessible points, but these would lie in  $\text{basin}(A)$  (or  $\text{basin}(B)$ ). This situation occurs in the previously mentioned pendulum example. Hence,  $S(R)$  might contain points on the basin boundary that are accessible from  $\text{basin}(A)$  (or  $\text{basin}(B)$ ) but not accessible from the transient set  $R \setminus S(R)$ . In example 2 below,  $S(R)$  contains such points in the basin boundary. Therefore, the accessible PIM triple procedure [NY2] for finding accessible points on  $S(R)$  is, generally speaking, not a procedure for finding accessible points on the basin boundary. We would like to point out that there are cases where  $S(R)$  equals the set  $(\text{basin boundary} \cap R)$ , (though this condition may be hard to verify). In such cases the ASST method (involving the accessible PIM triple procedure) might be used for finding accessible trajectories on the basin boundary.

**Example 2.** In this example, we illustrate the fact that  $S(R)$  can contain points that are not in the basin boundary, and for simplicity we present one-dimensional maps. Consider two one-dimensional maps with attractor  $A$  (which is  $-\infty$ ) and attractor  $B$  (which is  $+\infty$ ). Let  $f$  and  $g$  be the piecewise linear maps of which the graph is given in figure 2(a) and 2(b) respectively, such that  $g(y) = f(y)$  for all  $y \leq 1$ .



**Figure 2.** One-dimensional maps  $f$  and  $g$  (the graphs of  $f$  and  $g$  are given in 2(a) and 2(b) respectively). When we choose the transient region  $R$  to be the interval  $(-2, 3)$ , the stable set  $S(R; f)$  equals the basin boundary, and the stable set  $S(R; g)$  is strictly larger than the basin boundary.

Let  $p$  and  $q$  denote the two fixed points of  $g$  in  $(1, \infty)$ , and write  $m = \min\{g(y) : y \geq p\}$ . Assume  $1 < m < 2 < p < q < 3 < g(m)$ , see figure 2. The maps are constructed in such a way that  $\text{basin}\{A\}$  and  $\text{basin}\{B\}$  of  $g$  and  $f$  coincide. Hence, both maps have the same basin boundary and it is contained in the interval  $[-1, 1]$ . Note that the basin boundary is the set of all points in  $[-1, 1]$  that stay inside  $[-1, 1]$  under all positive iterates of the map  $f$  (or  $g$ ), and the basin boundary is fractal.

On the other hand we have, all points in  $(1, \infty)$  go to attractor  $B$  under forward iteration of the map  $f$ , whereas  $\text{basin}\{B\}$  for  $g$  includes a chaotic saddle in the open interval  $(2, 3)$ . When we choose the transient region  $R$  to be the open interval  $(-2, 3)$ , the stable sets  $S(R; f)$  and  $S(R; g)$  are the sets of points that stay in  $R$  under all forward iterates of  $f$  and  $g$  respectively. We have the basin boundary equals the stable set  $S(R; f)$ , but the stable set  $S(R; g)$  is strictly larger than the basin boundary. It can be shown that points of  $S(R; g) - S(R; f)$  can be found arbitrarily close to each point of the basin boundary.

We would like to address the following problem.

*Accessible basin boundary static restraint problem.* Given a segment  $J$  that has one end point in  $\text{basin}\{A\}$  and one end point in  $\text{basin}\{B\}$ , describe a procedure for finding a point on the basin boundary (in  $J \cap S(R)$ ) which is accessible from  $\text{basin}\{A\}$ .

We will state a procedure (the *accessible basin boundary refinement procedure*) for finding accessible points in  $M$  on the basin boundary. We will show it is valid (guaranteed to work) for the same class of hyperbolic dynamical systems as in [NY2], namely hyperbolic systems in which the unstable manifolds are one dimensional.

All the procedures are based on our presumed ability to specify an initial point  $p$  and compute the time  $T_R(p)$  its trajectory takes to escape from  $R$ . For applications, we need a 'dynamic' version of the 'static' problem above, since we want to produce numerical trajectories that are accessible from  $\text{basin}\{A\}$ . The 'dynamic' problem that is associated with the 'static' one is the following.

*Accessible basin boundary dynamic restraint problem.* Given a line segment  $J$  that has one end point in  $\text{basin}\{A\}$  and the other end point in  $\text{basin}\{B\}$ , describe a procedure for finding a numerical trajectory on the basin boundary that starts on  $J$  and which is accessible from  $\text{basin}\{A\}$ .

The ideas of the 'accessible basin boundary refinement procedure', which solved the 'static' problem, can be applied to solve the 'dynamic' problem, in such a way that implementation is possible on a computer. For more details, see the discussion in section 6.

The organization of the paper is as follows. In section 2 we present the 'accessible basin boundary refinement procedure'. Then, in section 3, we discuss some examples in which the straddle method involving this refinement procedure has been used. The main result for the validity of the refinement procedure for hyperbolic systems is stated precisely in section 4, and this result is proved in section 5. Section 6 is devoted to the discussion of the associated numerical method (the accessible basin boundary straddle trajectory method or ABST method) and related

numerical methods. Finally in section 7, the case of  $d$ -dimensional hyperbolic systems,  $d \geq 3$ , and smoothness of  $F$  are discussed.

## 2. The accessible basin boundary refinement procedure

Let the manifold  $M$ , the diffeomorphism  $F$ , the trident region  $R$ , and generalized attractors  $A$  and  $B$  be as before. Recall that we assume that each point that leaves  $R$  under iteration of  $F$  is either in  $\text{basin}(A)$  or in  $\text{basin}(B)$ . The *escape time*  $T_R(x)$  of a point  $x$  in  $R$  is defined by  $T_R(x) = \min\{n \geq 1 : F^n(x) \notin R\}$ , and  $T_R(x) = \infty$  if  $F^n(x) \in R$  for all  $n \geq 1$ . We say,  $T_R(x) = 0$  if  $x \notin R$ .

Let  $J$  be an unstable segment in  $R$ . The notation  $\{x, y\}$  for a pair means that  $x$  and  $y$  lie on  $J$ . Since  $J$  is homeomorphic to an interval, we may assume it has the ordering of an interval. For  $\{x, y\}$  we always assume for convenience that the ordering on  $J$  is such that we may write  $x < y$ , and denote  $[x, y]$  for the segment on  $J$  joining  $x$  and  $y$ . Let  $L \subset J$  be any connected subset of  $J$ . Assume  $L$  intersects the stable set  $S(R)$  transversally, and let  $\{a, b\}$  be a pair on  $L$ . For each  $\varepsilon > 0$ , an  $\varepsilon$ -refinement of  $\{a, b\}$  is a finite set of points  $a = g_0 < g_1 < \dots < g_N = b$  in  $[a, b]$ , such that

$$(\varepsilon/2) \cdot \rho([a, b], J) \leq \rho([g_k, g_{k+1}], J) \leq \varepsilon \cdot \rho([a, b], J)$$

for all  $k$ ,  $0 \leq k \leq N-1$ .

We say the pair  $\{a, b\}$  is a *straddle pair* if  $a \in \text{basin}(A)$  and  $b \in \text{basin}(B)$ . We call  $\{a, b\}$  a *proper straddle pair* if  $\{a, b\}$  is a straddle pair, and at least one of the points  $a$  and  $b$  is in the interior of  $L$ . If  $\{a, b\}$  is a (proper) straddle pair, then we call the interval  $[a, b]$ , a (proper) *straddle segment*. Our objective is to describe the 'accessible basin boundary refinement procedure' that selects in a unique way a proper straddle pair from any  $\varepsilon$ -refinement of a given straddle pair (on  $J$ ). When we repeatedly apply the procedure to the end points of the ever decreasing straddle segments (with lengths converging to zero), the resulting nested sequence converges to an accessible point  $p$  in the basin boundary; of course, this point  $p$  is in  $J \cap S(R)$ . The point  $p$  that we find is accessible using the curve  $[r, p]$ , for some  $r$  in  $J \cap \text{basin}(A)$ , so we say  $p$  is 'accessible from the left' ('accessible from  $\text{basin}(A)$ '), that is, from the side containing  $r$  (in  $\text{basin}(A)$ ). We could alternatively have chosen to approach from the right and we would expect to find a different point on the basin boundary. Since almost every point on  $J$  has finite escape time (see section 4), we can assume that all points of all refinements are chosen with finite escape time.

We now describe the accessible basin boundary refinement procedure which is the refinement procedure that generates a uniquely defined proper straddle pair from a given straddle pair. This procedure plays a dominant role in the method that generates a numerical trajectory on the basin boundary that is accessible from  $\text{basin}(A)$ . A slightly improved version is stated in section 4.

Let  $\{a, b\}$  be a straddle pair on a curve segment  $J$  such that  $a$  is contained in  $\text{basin}(A)$ , and  $b$  is contained in  $\text{basin}(B)$ . Let  $P = \{x_i : 0 \leq i \leq N(\varepsilon)\}$  be any  $\varepsilon/3$ -refinement of  $\{a, b\}$ , we of course have  $P \subset J$  and  $a = x_0 < x_1 < \dots < x_{N(\varepsilon)} = b$ . We choose the proper straddle pair  $\{a^*, b^*\}$  from  $P$  in the following way:

- (1) select  $b^*$  to be the leftmost point of  $P$  that is in  $\text{basin}(B)$ ;
- (2) define  $m$  to be the minimum of the escape time of the points in  $P$  to the left of  $b^*$ , and write  $a^0$  to denote the rightmost point to the left of  $b^*$  that has the minimum escape time  $m$ .

- (2a) If  $m < T_R(a)$  then choose  $a^* = a^0$ ; otherwise,  
 (2b) if  $m = T_R(a)$  then the choice of  $a^*$  depends on the grid  $P^*$  consisting of  $b^*$  and all the points in  $P$  to the left of  $b^*$  (that is,  $P^* = \{x \in P : x \in [a, b^*]\}$ ).  
 (i) If the grid  $P^*$  is not an  $\varepsilon$ -refinement of  $\{a, b^*\}$ , then choose  $a^* = a$ ; otherwise,  
 (ii) if the grid  $P^*$  is an  $\varepsilon$ -refinement of  $\{a, b^*\}$  then choose  $a^*$  to be the adjacent point in  $P^*$  to the right of  $a^0$ , unless  $b^*$  is that adjacent point, in which case choose  $a^* = a^0$ .

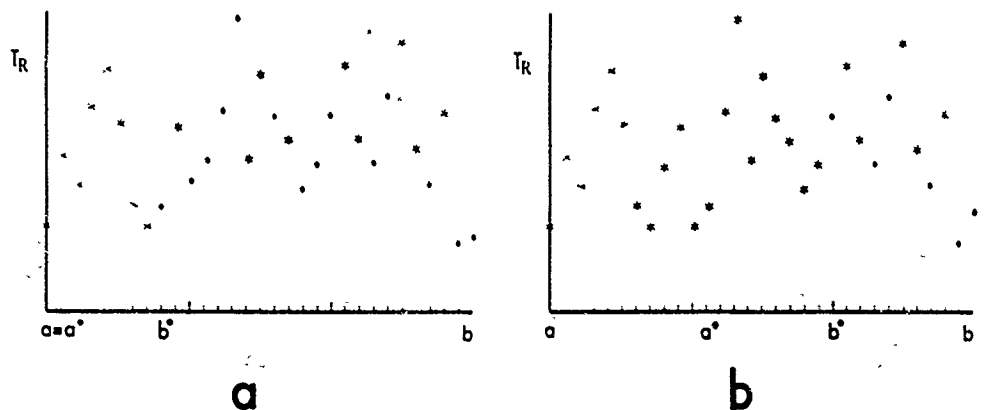
*Remark.* Assume that  $\varepsilon > 0$  is suitably chosen. In case of step (2b) the equality  $a^* = a^0$  does not occur and one has  $a^* > a^0$ .

(1) As the accessible basin boundary refinement procedure is applied repeatedly, step (2a) only occurs at most finitely many times, and the segment  $[a, a^*]$  in (2a) may include points that are in  $\text{basin}(B)$ . However, once step (2b) occurs, step (2a) will never occur again. When step (2b) is applied, the entire segment  $[a, a^*]$  (not just the grid points) is in  $\text{basin}(A)$  but  $[a, a^*]$  may include points that have escape time infinity. We would like to emphasize that all the points between  $a$  and  $a^*$  in step (2b) whose escape time is finite, go to attractor  $A$ . This is why the method produces an accessible point as the refinement is repeated. The problem of course is to find  $\varepsilon$  small enough.

(2) When  $a^*$  and  $b^*$  have been chosen, if the grid consisting of  $a^*$ ,  $b^*$  and all the points in  $P$  between  $a^*$  and  $b^*$  is still an  $\varepsilon$ -refinement of the pair  $\{a^*, b^*\}$ , then set  $a^* = a$  and  $b^* = b$  and apply step (2b). Repeat this until the grid  $\{x \in P : x \in [a^*, b^*]\}$  fails to be an  $\varepsilon$ -refinement of  $\{a^*, b^*\}$ . Notice that in cases when only step (2b) is repeated, the point  $b$  does not move.

(3) Under hypotheses in section 4, it is possible to repeatedly apply the accessible basin boundary refinement procedure obtaining a sequence of straddle pairs that converges to an accessible point on the basin boundary.

*Example 3.* The purpose of this example is to illustrate the accessible basin boundary refinement procedure in a graphical way. We choose  $\varepsilon = 0.1$ . Let  $\{a, b\}$  be a straddle pair, and let  $P$  be an  $\varepsilon/3$ -refinement of  $\{a, b\}$ . We assume that  $P$  is on



**Figure 3.** The accessible basin boundary refinement procedure. In figure 3(a) the grid on  $[a, b^*]$  is not an  $\varepsilon$ -refinement of  $\{a, b^*\}$  and so  $a$  does not move; in figure 3(b) the grid on  $[a, b^*]$  is an  $\varepsilon$ -refinement of  $\{a, b^*\}$  and so  $a$  moves to the right.

the straight line segment that joins  $a$  with  $b$  and that the grid points are equally spaced, so  $P$  consists of 31 grid points. In figure 3 the escape time of a grid point  $x$  in  $P$  is represented by a star if  $x$  is in  $\text{basin}\{A\}$ , and it is represented by a dot if  $x$  is in  $\text{basin}\{B\}$ .

In figure 3(a) we have  $b^* = x_8$ . The grid  $P^* = \{x \in P : x \in [a, b^*]\}$  is not an  $\varepsilon$ -refinement of  $\{a, b^*\}$ , since the distance between two adjacent points equals  $\|b^* - a\|/8$  which is greater than  $\varepsilon \cdot \|b^* - a\|$ . Hence, we choose  $a^* = a$ . In figure 3(b) we have  $b^* = x_{20}$ . The grid  $P^* = \{x \in P : x \in [a, b^*]\}$  is an  $\varepsilon$ -refinement of  $\{a, b^*\}$ , since the distance between two adjacent points equals  $\|b^* - a\|/20$  which is smaller than  $\varepsilon \cdot \|b^* - a\|$ . Since  $T_R(x_0) = T_R(x_8) = T_R(x_{10}) = m$ , we choose  $a^* = x_{11}$  as indicated in the figure.

### 3. Applications

The objective of the paper is to present the accessible basin boundary refinement procedure which enables us to obtain accessible numerical trajectories on the basin boundary. We also prove that this numerical procedure works in ideal cases. While we believe that the hyperbolicity hypotheses (stated in section 4) are often satisfied, they are nonetheless in practice difficult or impossible to verify. While chaotic attractors are usually not hyperbolic, the sets we look at are not attractors. We do observe that frequently we can successfully use the procedure to obtain pictures of the accessible points on the basin boundary.

In all the examples below, the pictures were obtained by using the Dynamics Program [Y]. In these pictures,  $\text{basin}\{X\}$  is obtained as follows: for a  $960 \times 544$  grid, use each grid point as initial value and assign to each grid point a colour (respectively, no colour) if its trajectory converges to  $X$  (respectively, stays away from  $X$ ). The set of coloured grid points is in  $\text{basin}\{X\}$ , and the non-coloured grid points are outside  $\text{basin}\{X\}$ . In all the pictures for which one of the numerical procedures has been applied in order to produce a single numerical trajectory, have been obtained by selecting  $\varepsilon = 1/30$  as default value (see also section 6).

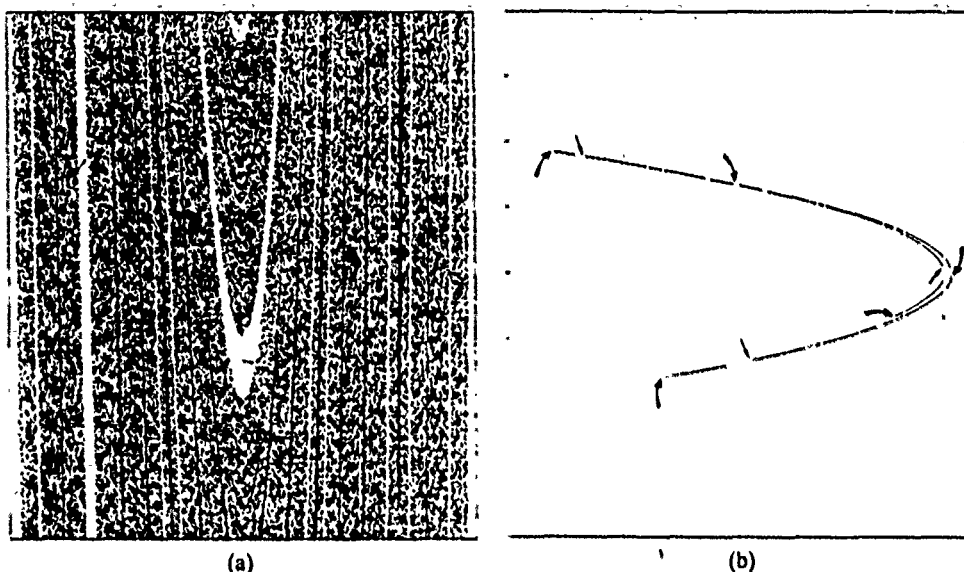
#### 3.1. Hénon map

Let the diffeomorphism  $F$  acting on the plane be given by

$$F(x, y) = (\rho - x^2 + \mu \cdot y, x).$$

The map  $F$  is equivalent under a change of variables to the Hénon map  $(1 - \rho \cdot X^2 + Y, \mu \cdot X)$ . For a first example, we choose the parameters  $\rho = 1.812\,579\,70$  and  $\mu = 0.022\,864\,30$ ; these parameters are due to Grassberger and Cvitanović (personal communication). For these parameters attracting cycles with period 3 and period 5 coexist. Let  $D_1$  and  $D_2$  be closed balls of radius 0.01 centred at one of the points of the attracting period 3 cycle and 5 cycle respectively. We choose the transient region  $R$  to be the open set  $\{(x, y) : -2 < x < 2, -4 < y < 4\}$  minus the closed balls  $D_1$  and  $D_2$ .

Let  $A$  and  $B$  be the attractors with period 3 and period 5 respectively. The white area in figure 4(a) is  $\text{basin}\{A\}$ ; the black area is  $\text{basin}\{B\}$ . By using the bisection procedure (see also section 6), we obtain a straddle trajectory (that is, a numerical



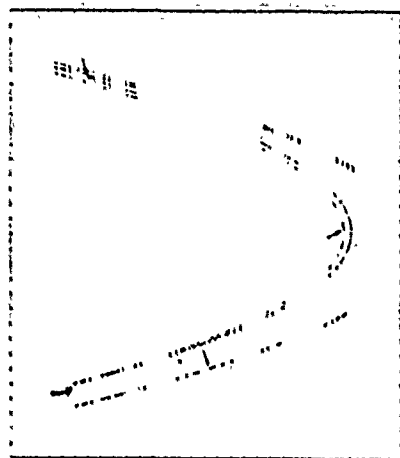
**Figure 4.** (a) The white area is basin(A) and includes the period 3 attractor, the black area is basin(B) and includes the period 5 attractor in the region  $-2 < x < 2$ ,  $-4 < y < 4$  of the Hénon map with parameter values  $\rho = 1.812\,579\,70$ ,  $\mu = 0.022\,864\,30$ . (b) Straddle trajectory using the bisection procedure for the Hénon map ( $\rho = 1.812\,579\,70$ ,  $\mu = 0.022\,864\,30$ ) in the transient region  $\{(x, y) : -2 < x < 2, -4 < y < 4\}$  minus two closed balls of radius 0.01 centred at a point of each attractor. The three saddle periodic points on the basin boundary that are accessible from basin(A) and the five saddle periodic points on the basin boundary that are accessible from basin(B) are indicated by straight and curved arrows respectively.

trajectory) on the basin boundary consisting of more than 100 000 points (actually tiny intervals); the result is presented in figure 4(b).

By using the accessible basin boundary refinement procedure we obtain a period 3 saddle when the left point  $a$  is chosen in basin(A), and a period 5 saddle when the left point  $a$  is chosen in basin(B). The accessible period 3 and period 5 saddles on the chaotic saddle are indicated by arrows in figure 4(b). Therefore, the set of all points accessible from basin(A) are the stable manifolds of the points of the period 3 saddle, and all points accessible from basin(B) are the stable manifolds of the points of the period 5 saddle.

For a second example, we select the values  $\rho = 2.66$ ,  $\mu = 0.3$ . The map  $F$  has two attractors  $A$  and  $B$ , where  $A$  and  $B$  denote the attractors infinity and a cycle with period 3 respectively. The box  $\{(x, y) : -3 < x < 3, -3 < y < 3\}$  contains a chaotic saddle, and we select the transient region  $R$  to be the open set  $\{(x, y) : -3 < x < 3, -3 < y < 3\}$  minus the ball of radius 0.005 centred at a point of attractor  $B$ . Using the bisection procedure results in one numerical trajectory, that has been presented in figure 5.

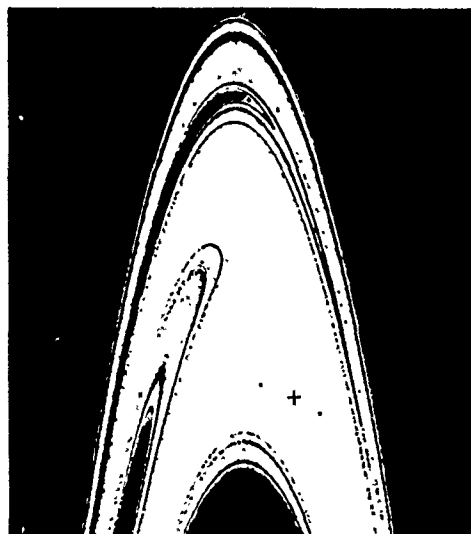
By using the accessible basin boundary refinement procedure we obtain a period 1 saddle when the left point  $a$  is chosen in basin(A), and a period 3 saddle when the left point  $a$  is chosen in basin(B). The accessible period 1 and period 3 saddles on the chaotic saddle are indicated by arrows in figure 5. So, the set of all points accessible from basin(A) is the stable manifold of the period 1 saddle, and the set of



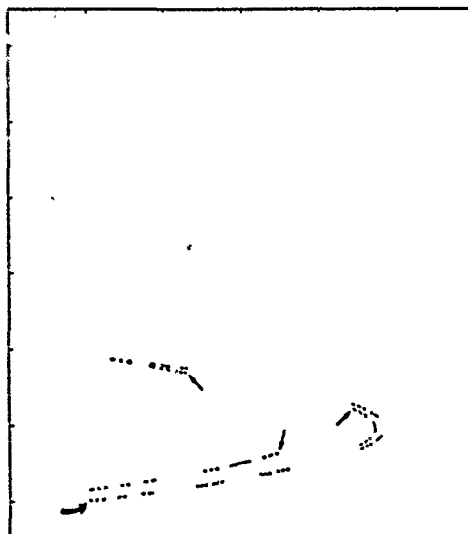
**Figure 5.** Straddle trajectory using the bisection procedure for the Hénon map ( $\rho = 2.66$ ,  $\mu = 0.3$ ) in the transient region  $\{(x, y) : -3 < x < 3, -3 < y < 3\}$  minus a closed ball of radius 0.005 centred at a point of attractor  $B$  (the period 3 attractor). The fixed point on the basin boundary that is accessible from basin( $A$ ) (where  $A = \infty$ ), and the three saddle periodic points on the basin boundary that are accessible from basin( $B$ ) are indicated by curved and straight arrows respectively.

all points accessible from basin( $B$ ) are the stable manifolds of the points of the period 3 saddle.

For a third example of this map, we select the parameter values  $\rho = 1.405$ ,  $\mu = -0.3$ . The map  $F$  has two coexisting attractors, namely, a period 2 cycle (attractor  $A$ ) and the attractor infinity (attractor  $B$ ). The box  $\{(x, y) : -3 < x < 3, -3 < y < 11\}$  contains a chaotic saddle. Basin( $A$ ) is the white area in figure 6(a) (the two points of attractor  $A$  are marked by a dot in the figure), and basin( $B$ ) is black in figure 6(a).



(a)



(b)

**Figure 6.** (a) The white area is basin( $A$ ) and includes the period 2 attractor, the black area is basin( $B$ ) (where  $B = \infty$ ) in the region  $\{(x, y) : -3 < x < 3, -3 < y < 11\}$  of the Hénon map with parameter values  $\rho = 1.405$ ,  $\mu = -0.3$ . Attractor  $A$  is marked by two dots, and a saddle fixed point in basin( $A$ ) is marked by a cross. (b) Straddle trajectory using the bisection procedure for the Hénon map ( $\rho = 1.405$ ,  $\mu = -0.3$ ) in the transient region  $\{(x, y) : -3 < x < 3, -3 < y < 11\}$  minus a closed ball of radius 0.2 centred at a point of attractor  $A$ . The three saddle periodic points on the basin boundary that are accessible from basin( $A$ ) and the saddle fixed point on the basin boundary that is accessible from basin( $B$ ) are indicated by straight and curved arrows respectively.

We select the transient region  $R$  to be the open set  $\{(x, y): -3 < x < 3, -3 < y < 11\}$  minus the ball of radius 0.2 centred at a point of attractor  $A$ . Using the bisection procedure results in one numerical trajectory, that has been presented in figure 6(b). The PIM triple procedure may result in a saddle fixed point that is in  $\text{basin}(A)$ ; this saddle point is marked by a cross in figure 6(a). If we select the transient region to be the region  $R$  minus a ball of radius 0.2 centred at this saddle fixed point, then applying the PIM triple procedure results a similar numerical trajectory as in figure 6(b). Notice that the ball including the saddle fixed point is in  $\text{basin}(A)$ .

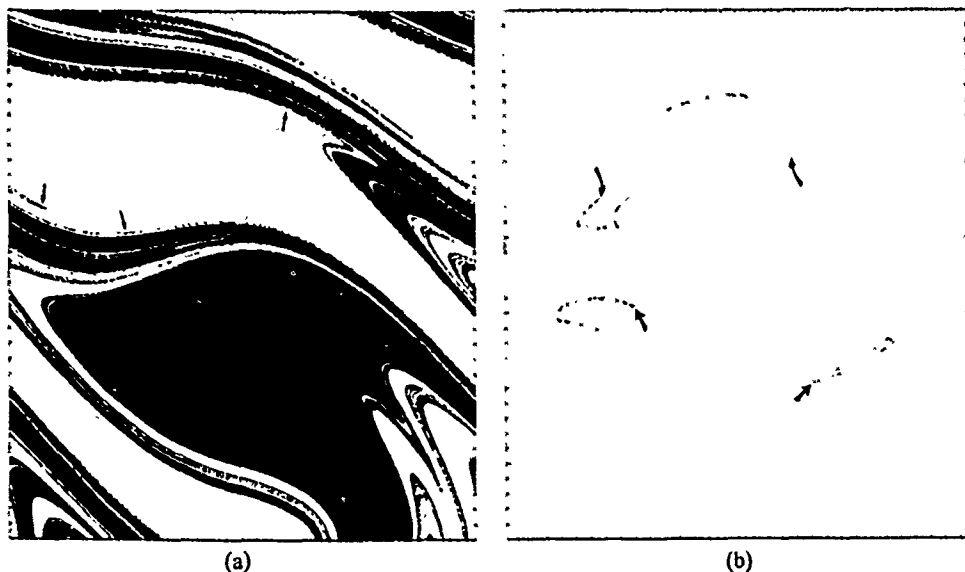
By using the accessible basin boundary refinement procedure we obtain a period 3 saddle, when the left point  $a$  is chosen in  $\text{basin}(A)$ , and a period 1 saddle, when the left point  $a$  is chosen in  $\text{basin}(B)$ . The points of the accessible period 3 saddle on the chaotic saddle are indicated by arrows in figure 6(a). So, the set of all points accessible from  $\text{basin}(A)$  are the stable manifolds of the points of the period 3 saddle, and the set of all points accessible from  $\text{basin}(B)$  is the stable manifold of the period 1 saddle.

Note that the invariant set of points in the transient region consists of at least three basic sets, namely, (1) the period 2 attractor, (2) the saddle fixed point in  $\text{basin}(A)$  and (3) the chaotic saddle on the basin boundary.

### 3.2. Pendulum

We consider the differential equation

$$x''(t) + vx'(t) + \sin x(t) = f \cos(t).$$



**Figure 7.** (a) The white area is  $\text{basin}(A)$  and the black area is  $\text{basin}(B)$  (where  $A = (-0.472615, 2.037084)$  and  $B = (-0.478014, -0.608233)$  are fixed point attractors) in the region  $\{(x, y): -\pi < x < \pi, -3 < y < 4\}$  of the time- $2\pi$  map of the forced pendulum  $x''(t) + 0.2x'(t) + \sin x(t) = 2 \cos(t)$ . The three saddle periodic points on the basin boundary that are accessible from  $\text{basin}(A)$  are indicated by arrows. (b) Two straddle trajectories using the PIM triple refinement procedure for the time- $2\pi$  map of  $x''(t) + 0.2x'(t) + \sin x(t) = 2 \cos(t)$  in the transient region  $\{(x, y): -\pi < x < \pi, -3 < y < 4\}$  minus two closed balls of radius 0.05 centred at the fixed point attractors  $A$  and  $B$ , one trajectory in both  $\text{basin}(A)$  and  $\text{basin}(B)$ . The two saddle periodic 2 orbits on the stable set that are accessible from the transient set  $R \setminus S(R)$  are indicated by arrows.

We choose the parameter values  $v = 0.2$  and  $f = 2$ . For these parameters, the time- $2\pi$  map has two stable fixed points  $A$  and  $B$ . In figure 7(a), basin $\{A\}$  is coloured white and basin $\{B\}$  is coloured black. It was already observed [GOY2] that there was transient behaviour in the basin $\{A\}$  and basin $\{B\}$ . We choose the transient region to be the rectangle  $\{(x, y) : -\pi < x < \pi, -3 < y < 4\}$  minus two balls (of radius 0.05) centred at the attractors  $A$  and  $B$ . By using the PIM triple procedure for two different transient regions, we obtain two numerical trajectories. The result for the choice of the interval with end points  $(-3, -3)$  and  $(3, 4)$  is a trajectory lying in basin $\{A\}$ ; and the segment from  $(-3, 4)$  to  $(3, -3)$  results in a numerical trajectory lying in basin $\{B\}$ . Both trajectories are presented in figure 7(b).

By using the accessible PIM triple procedure we obtain period 2 saddles, see also the discussion in section 6. The result for the segment from  $(-3, -3)$  to  $(3, 4)$  is a period 2 saddle on the chaotic saddle in basin $\{A\}$ , and the segment from  $(-3, 4)$  to  $(3, -3)$  results in a period 2 saddle on the chaotic saddle in basin $\{B\}$ . The points of these accessible period 2 saddles on the chaotic saddle are indicated by arrows in figure 7(b). The set of all accessible points on the two chaotic saddles are the stable manifolds of the points of these period 2 saddles.

By using the accessible basin boundary refinement procedure we obtain two period 3 saddles: one is accessible from basin $\{A\}$ , and the other one is accessible from basin $\{B\}$ . The points of the period 3 saddle that is accessible from basin $\{A\}$  and is on the basin boundary, are indicated by arrows in figure 7(a). The set of all points on the basin boundary that are accessible from basin $\{A\}$ , are the stable manifolds of the points of this period saddle. A similar result as above holds for the points on the basin boundary that are accessible from basin $\{B\}$ .

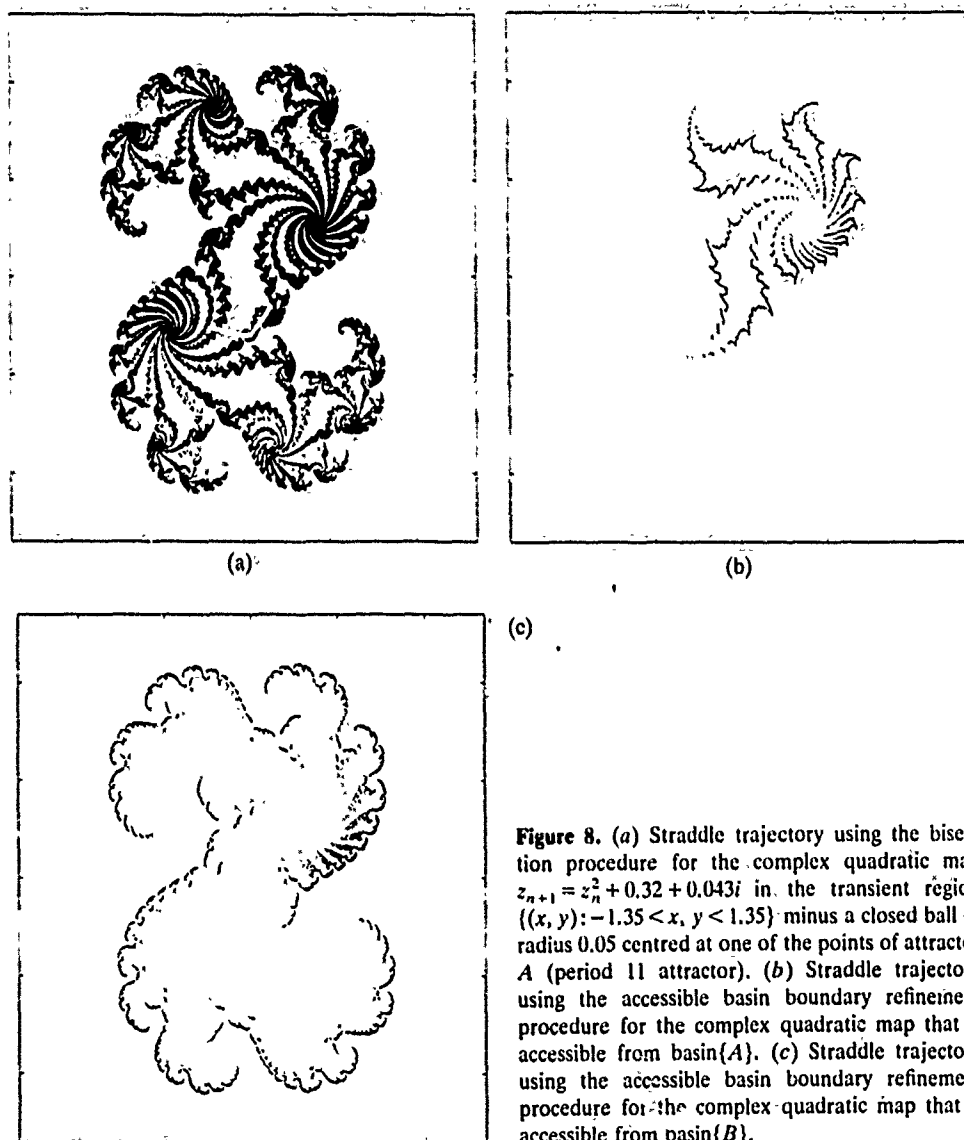
### 3.3. Complex quadratic map

We consider the quadratic map in the complex plane given by

$$z_{n+1} = z_n^2 + 0.32 + 0.043i.$$

For this system two attractors coexist, namely, a period 11 attractor (attractor  $A$ ) and the attractor infinity (attractor  $B$ ). Let  $D$  be a closed ball of radius 0.05 centred at a point of attractor  $A$ . We choose the transient region  $R$  to be the open set  $\{(x, y) : -1.35 < x < 1.35, -1.35 < y < 1.35\}$  minus the ball  $D$ . The basin boundary straddle trajectory resulting from the bisection procedure is presented in figure 8(a). The accessible basin boundary straddle trajectory resulting from the accessible basin boundary refinement procedure, a trajectory of which all the points are accessible from basin $\{A\}$  is presented in figure 8(b), and the accessible basin boundary straddle trajectory of which all the points are accessible from basin $\{B\}$  is presented in figure 8(c).

The choice of this equation was motivated by the picture of the Julia set in [PR]. The reader should compare our figure 8(a) with figure 25 in [PR]. We would like to point out that the basin boundary of this system (the Julia set) is two dimensionally unstable; thus our results are not valid for this example.



**Figure 8.** (a) Straddle trajectory using the bisection procedure for the complex quadratic map  $z_{n+1} = z_n^2 + 0.32 + 0.043i$  in the transient region  $\{(x, y) : -1.35 < x, y < 1.35\}$  minus a closed ball of radius 0.05 centred at one of the points of attractor  $A$  (period 11 attractor). (b) Straddle trajectory using the accessible basin boundary refinement procedure for the complex quadratic map that is accessible from basin  $\{A\}$ . (c) Straddle trajectory using the accessible basin boundary refinement procedure for the complex quadratic map that is accessible from basin  $\{B\}$ .

#### 4. Results

In section 2 we presented the accessible basin boundary refinement procedure for finding a point on the basin boundary in the transient region, which is accessible from basin  $\{A\}$ . First, we formulate a refinement procedure which is a slightly improved version of the accessible basin boundary refinement procedure.

We will describe inductively how to refine our proper straddle pairs. Given a straddle pair  $\{a_n, b_n\}$ , we have  $a_n$  is contained in basin  $\{A\}$ , and  $b_n$  is contained in basin  $\{B\}$ . Given any  $\varepsilon/3$ -refinement  $P_n = \{x_i : 0 \leq i \leq N(\varepsilon)\}$  of  $\{a_n, b_n\}$ , we of course have  $a_n = x_0 < x_1 < \dots < x_{N(\varepsilon)} = b_n$ . We choose the next proper straddle pair  $\{a_{n+1}, b_{n+1}\}$  from  $P_n$  in the following way.

- (1) Select  $b_{n+1}$  to be the leftmost point of  $P_n \cap \text{basin}\{B\}$ .

(2) Define  $m_n = \min\{T_R(x) : x \in P_n \text{ and } x < b_{n+1}\}$ ;

$$a_{n+1}^0 = \max\{x \in Q_n : x < b_{n+1} \text{ and } T_R(x) = m_n\}.$$

(2a) If  $m_n < T_R(a_n)$  then choose  $a_{n+1} = a_{n+1}^0$ ; otherwise,

(2b) If  $m_n = T_R(a_n)$  then in order to choose  $a_{n+1}$  we write

$$Q_n = \{x \in P_n : \dots [a_n, b_{n+1}]\}.$$

$$a_{n+1}^+ = \text{minimum of the set } \{x \in Q_n : a_{n+1}^0 < x < b_{n+1}\},$$

unless this set is empty in which case  $a_{n+1}^+ = a_{n+1}^0$ ;

$$a_{n+1}^1 = \max\{x \in Q_n : x < b_{n+1} \text{ and } T_R(x) = T_R(a_{n+1}^+)\}.$$

Case (i) If  $Q_n$  is not an  $\varepsilon$ -refinement of  $\{a_n, b_{n+1}\}$ , then choose  $a_{n+1} = a_n$ ; otherwise,

Case (ii) If  $Q_n$  is an  $\varepsilon$ -refinement of  $\{a_n, b_{n+1}\}$  then choose  $a_{n+1} = a_{n+1}^1$ .

*Remark* (1) For the convenience of the reader, if  $\varepsilon > 0$  is chosen suitably, then  $a_n \leq a_{n+1}^0 < a_{n+1}^+ \leq a_{n+1}^1 < b_{n+1}$  and  $m_n = T_R(a_{n+1}^0) < T_R(a_{n+1}^+) = T_R(a_{n+1}^1)$ . Note that  $Q_n$  might fail to be an  $\varepsilon$ -refinement of  $\{a_n, b_{n+1}\}$  in that the distance between some pair of consecutive points in  $Q_n$  might be bigger than  $\varepsilon \cdot \rho([a_n, b_{n+1}]_J)$ .

(2) Under the hypotheses below it is possible to repeatedly apply the improved refinement procedure above obtaining a sequence  $\{\{a_n, b_n\}\}_{n \geq 0}$  that settles down to an accessible point on the basin boundary.

In the description of the refinement procedure above, we assumed that there exists an  $\varepsilon > 0$  for which every  $\varepsilon$ -refinement of a straddle pair  $\{a_n, b_n\}$  includes a proper straddle pair  $\{a_{n+1}, b_{n+1}\}$  such that  $[a_n, a_{n+1}]_J$  is in  $\text{basin}\{A\}$ , and the length of the straddle segment  $[a_{n+1}, b_{n+1}]_J$  is at most  $(1 - \varepsilon/2)$  times the length of the previous straddle segment  $[a_n, b_n]_J$ . We will justify these concepts.

Let the manifold  $M$  and the diffeomorphism  $F$  be as in the introduction. We assume that  $A$  and  $B$  are two generalized attractors such that each attractor is contained either in  $A$  or in  $B$ . Recall that a subset  $\Lambda$  of  $M$  is *hyperbolic* if it is closed and  $F$ -invariant and the tangent bundle  $T_\Lambda M$  splits into  $dF$ -invariant sub-bundles  $E^s$  and  $E^u$  on which  $dF$  is uniformly contracting and uniformly expanding respectively. A hyperbolic set  $\Lambda$  is called *saddle-hyperbolic* if  $\dim E^s \geq 1$  and  $\dim E^u \geq 1$ . In [NY2] we defined a region  $R$  to be a *saddle-hyperbolic transient region* if  $R$  satisfies all the following conditions:

(A1)  $R$  is a transient region;

(A2) *hyperbolicity property*:  $\text{inv}(R)$  is a non-empty saddle-hyperbolic set;

(A3) *boundary property*:  $\bar{U}(R) \cap \partial R$  is mapped outside the closure  $\bar{R}$  of  $R$ ;

(A4) *intersection property*: each non-trivial component  $\gamma$  of  $U(R)$  is an unstable segment, that is,  $\gamma$  intersects  $\text{Inv}(R)$ ; note that such a segment  $\gamma$  must intersect  $S(R)$  transversally.

In this paper, we say a transient region  $R$  satisfies the *basin boundary property* if (1) each point in  $R \setminus S(R)$  is contained in either  $\text{basin}\{A\}$  or  $\text{basin}\{B\}$ , (2) the sets  $R \cap \text{basin}\{A\}$  and  $R \cap \text{basin}\{B\}$  are non-empty, and (3) the  $R \cap \text{basin}$  boundary is positively invariant (that is,  $F$  maps the basin boundary into itself). We define a region  $R$  to be a *basin boundary transient region* if  $R$  is a saddle-hyperbolic transient region and  $R$  satisfies the basin boundary property.

For a basin boundary transient region  $R$ , and  $\varepsilon > 0$ , the properties (A1) and (A2) imply that the escape time of almost every point on an unstable segment is finite. (A result due to Bowen and Ruelle [BR] shows that  $S(R)$  has Lebesgue measure zero.) Hence, one may assume that such a refinement does not intersect the stable set  $S(R)$ . The basin boundary property implies that each point that escapes from  $R$  under iteration of the map  $F$  is either in  $\text{basin}(A)$  or in  $\text{basin}(B)$ .

If  $R$  is a basin boundary transient region, then the escape time map  $T_R$  restricted to an unstable segment  $J \subset U(R)$  has the following two properties (see [NY2]).

(i) All the points in a chosen segment  $[a, b]_J$  on  $J$  will escape from  $R$  if and only if no  $\varepsilon$ -refinement of  $\{a, b\}$  includes a PIM triple (that is, a triple  $(p, r, q)$  on  $J$  such that  $T_R(r) > T_R(p)$ ,  $T_R(r) > T_R(q)$ , and  $\rho([p, q]_J) < \rho([a, b]_J)$ ).

(ii)  $T_R$  is locally constant on an open subset of full measure of  $J$ , and if  $T_R(x) < \infty$  and  $x$  is a point of discontinuity of  $T_R$  then  $\liminf_{y \rightarrow x} T_R(y) = T_R(x)$  and  $\limsup_{y \rightarrow x} T_R(y) = T_R(x) + 1$ .

We assume throughout that  $\dim E^u = 1$ . For the sake of simplicity, we assume that  $d = 2$ ; the more difficult case  $d \geq 3$  will be discussed in section 7.

From now on, we will assume that  $R$  is a basin boundary transient region for  $F$ , and that  $J \subset U(R)$  denotes an unstable segment. The proof of the proposition below, will follow immediately from the propositions 5.1 and 5.2.

**Proposition.** There exists a finite set of periodic points  $P^u$  in  $\text{Inv}(R)$  such that (1) each point in  $P^u$  is accessible from  $R \setminus S(R)$ , and (2) for  $x \in S(R)$ , the point  $x$  is accessible from  $R \setminus S(R)$  if and only if  $x \in W^s(p)$  for some  $p \in P^u$ .

**Corollary.** Each accessible point on the basin boundary is in the stable manifold of some periodic point.

Since  $J$  is an unstable segment, recall that this implies that both ends of  $J$  are in the boundary of the transient region  $R$ . We know by the intersection assumption that  $J$  intersects the stable set  $S(R)$ . Obviously, if  $\{a, b\}$  is a straddle pair, then there exist proper straddle pairs in every  $\varepsilon$ -refinement of  $\{a, b\}$ , for each  $\varepsilon$ ,  $0 < \varepsilon \leq 0.5$ .

The next result deals with the convergence of the sequence of nested proper straddle segments  $[a_{n+1}, b_{n+1}]_J \subset [a_n, b_n]_J$  on  $J$ . A sequence of straddle segments  $\{[a_n, b_n]_J\}_{n \geq 0}$  on  $J$  is called a *straddle segment sequence* if  $\{a_{n+1}, b_{n+1}\}$  is in an  $\varepsilon$ -refinement of the straddle pair  $\{a_n, b_n\}$  for all  $n$ . We say  $\{[a_n, b_n]_J\}_{n \geq 0}$  is the *accessible straddle segment sequence* if  $\{a_n, b_n\}$  is selected using the accessible basin boundary refinement procedure for all  $n$ . For every  $\varepsilon$ ,  $0 < \varepsilon \leq 0.5$ , each straddle segment sequence  $\{[a_n, b_n]_J\}_{n \geq 0}$  converges to a point on the basin boundary. In section 5 we will show that there exists  $\varepsilon > 0$  (depending on  $F$  and  $R$ ) such that for every accessible straddle segment sequence  $\{[a_n, b_n]_J\}_{n \geq 0}$  there is an integer  $N \geq 0$  such that for every integer  $n \geq N$  the straddle segment  $[a_n, a_{n+1}]_J$  is contained in  $\text{basin}(A)$ . This number  $\varepsilon$  also appears in the result stated below. The main result stated below implies that the accessible basin boundary refinement procedure is valid.

**Theorem.** There exists  $\varepsilon > 0$  (depending on  $F$  and  $R$ ) such that every accessible straddle segment sequence converges to an accessible point on the basin boundary.

## 5. Proofs

### 5.1. Preliminaries

Let the manifold  $M$ , the distance  $\rho$  on  $M$ , and diffeomorphism  $F$  be as before. We assume that  $R$  is a basin boundary transient region for the diffeomorphism  $F$ , and that there are generalized attractors  $A$  and  $B$  such that each point that eventually leaves  $R$  is either in  $\text{basin}\{A\}$  or in  $\text{basin}\{B\}$ . Recall that the non-wandering set  $\Omega$  (that is, the set of all points  $x$  in  $M$  such that for every open neighbourhood  $V$  of  $x$  there exists  $n \geq 1$  for which  $F^n(V) \cap V$  is non-empty) can uniquely be decomposed into a finite collection of disjoint closed invariant subsets and on each of these subsets  $F$  has a dense orbit; these maximal invariant subsets of  $\Omega$  appearing in the decomposition are called the *basic sets* (see e.g. [GH] for the definitions and several properties of uniformly hyperbolic systems). From now on, let  $\Gamma$  denote a basic set of  $F$ . From the definition of  $\text{Inv}(R)$  it follows immediately that either  $\Gamma \subset \text{Inv}(R)$  or  $\Gamma \cap \text{Inv}(R)$  is empty. Thus, we can decompose  $\text{Inv}(R)$  into finitely many basic sets. Note that ' $\Gamma \cap \text{Inv}(R)$  is empty' does not imply ' $\Gamma \cap R$  is empty', and ' $\Gamma \cap R$  is non-empty' does not imply ' $\Gamma \cap \text{Inv}(R)$  is non-empty'.

Recall that for  $z \in \Omega$  the stable manifold  $W^s(z)$  of  $z$  is the set of points  $x$  for which  $\rho(F^n(z), F^n(x)) \rightarrow 0$  as  $n \rightarrow \infty$ , and the unstable manifold  $W^u(z)$  of  $z$  is the set of points  $x$  for which  $\rho(F^{-n}(z), F^{-n}(x)) \rightarrow 0$  as  $n \rightarrow \infty$ . The local stable manifold  $W_{\text{loc}}^s(z)$  of  $z$  (of size  $\beta$ ) is the set of points  $x$  in  $W^s(z)$  such that  $\rho(F^n(z), F^n(x)) \leq \beta$  for all integers  $n \geq 0$ , and the local unstable manifold  $W_{\text{loc}}^u(z)$  of  $z$  is the set of points  $x$  in  $W^u(z)$  such that  $\rho(F^{-n}(z), F^{-n}(x)) \leq \beta$  for all  $n \geq 0$ , where  $\beta > 0$ . When the stable or unstable manifold is a curve, we write  $W_{\text{loc}}^{\sigma+}(z)$  and  $W_{\text{loc}}^{\sigma-}(z)$  for the two components of  $W_{\text{loc}}^{\sigma}(z) \setminus \{z\}$ , where  $\sigma$  is either  $s$  or  $u$ .

We call  $\Gamma$  a *trivial* basic set if  $\Gamma$  consists of one periodic orbit, and we call  $\Gamma$  a *non-trivial* basic set if  $\Gamma$  includes more than one periodic orbit. Assume that  $\Gamma$  is non-trivial; we call  $\Gamma$  *periodic* if there exists  $m \in \mathbb{N}$  such that  $F^m$  has no dense orbit on  $\Gamma$ , and we call  $\Gamma$  *non-periodic* if it is not periodic.

We will see below that the structure of  $\text{Inv}(R)$  is essentially controlled by finite sets of periodic points. Recall that  $x$  in  $\text{Inv}(R)$  is accessible from an open set  $V$  if there is a curve  $\gamma$  such that  $\gamma \setminus \{x\}$  lies in  $V$ . If we choose  $V$  to be the transient set  $R \setminus S(R)$ , and if  $x$  in  $\text{Inv}(R)$  is accessible from  $R \setminus S(R)$  it is always possible to choose this curve  $\gamma$  to be a piece of the unstable manifold  $W^u(x)$ , that is,  $\gamma$  can be chosen to be either  $W_{\text{loc}}^{u+}(x)$  or  $W_{\text{loc}}^{u-}(x)$ . Notice if  $x$  is accessible from  $R \setminus S(R)$  and  $\gamma = W_{\text{loc}}^{u+}(x)$ , then  $x$  is not a limit point of  $W_{\text{loc}}^{u+}(x) \cap \Omega$ . Similarly, if we choose  $V$  to be the open set  $R \setminus U(R)$ , and if  $x$  in  $\text{Inv}(R)$  is accessible from  $R \setminus U(R)$  it is always possible to choose this curve  $\gamma$  to be a piece of the stable manifold  $W^s(x)$ , that is,  $\gamma$  can be chosen to be either  $W_{\text{loc}}^{s+}(x)$  or  $W_{\text{loc}}^{s-}(x)$ . Applying a result due to Newhouse and Palis [NP], we obtain the following.

**Proposition 5.1.** There exists a finite set  $P$  of periodic points in  $\text{Inv}(R)$ ,  $P = P^u \cup P^s$ , such that each point in  $\text{Inv}(R)$  that is accessible from  $R \setminus S(R)$  is in  $W^s(p)$  for some  $p$  in  $P^u$ , and each point in  $\text{Inv}(R)$  that is accessible from  $R \setminus U(R)$  is in  $W^u(p)$  for some  $p$  in  $P^s$ .

*Proof.* For a proof, see Newhouse and Palis [NP].  $\square$

Palis and Takens [PT] have shown that there exist regions in  $M$ , whose boundaries are segments in the stable and unstable manifolds of these finite sets of periodic points  $P^s$  and  $P^u$ , such that the intersection of the union of these regions with the saddle basic set  $\Gamma$  is a Markov partition for  $\Gamma$ , see Bowen [B] for the notion of Markov partition.

**Proposition 5.2.** Assume  $\Gamma$  is a non-trivial non-periodic basic set in  $\text{Inv}(R)$ , and let  $z \in \Gamma$  be fixed. Let  $P^s$  and  $P^u$  be as above. There exist finitely many disjoint regions  $R_i$  being diffeomorphic images of the square  $B = [-1, 1] \times [-1, 1]$ , say  $R_i = g_i(B)$ ,  $1 \leq i \leq N$  for some  $N \in \mathbb{N}$ , and a connected subset  $I^u$  of  $W^u(z)$  such that:

- (1)  $\Gamma \cap R_i$  is non-empty for all  $i$ ;
- (2)  $\Gamma \subset \bigcup_{i=1}^N R_i$ ;
- (3)  $F(\partial_s R_i) \subset \bigcup_{j=1}^N \partial_s R_j$  and  $F^{-1}(\partial_u R_i) \subset \bigcup_{j=1}^N \partial_u R_j$ , where  $\partial_s R_i = g_i(\{(x, y) : |x| = 1, |y| \leq 1\})$  and  $\partial_u R_i = g_i(\{(x, y) : |x| \leq 1, |y| = 1\})$  are connected subsets in the stable set  $W^s(P^u \cap \Gamma)$  and the unstable set  $W^u(P^s \cap \Gamma)$  respectively; and
- (4) for every  $i$ ,  $I^u \cap R_i$  consists of exactly one component and  $\partial(I^u \cap R_i) \subset \bigcup_{j=1}^N \partial_s R_j$ ,  $1 \leq i \leq N$ .

*Proof.* For a proof, see Palis and Takens [PT].  $\square$

Recall that  $R$  is a basin boundary transient region, and  $\Gamma$  a basic set in  $\text{Inv}(R)$ . From now on, let the point  $z \in \Gamma$ , the regions  $R_i$ ,  $1 \leq i \leq N$ , and the segment  $I^u \subset W^u(z)$  be as in proposition 5.2. There exist a  $C^{1+\alpha}$  stable foliation  $\mathcal{F}^s$  on a neighbourhood  $V_\Gamma^s$  of  $\Gamma$  and a  $C^{1+\alpha}$  unstable foliation  $\mathcal{F}^u$  on a neighbourhood  $V_\Gamma^u$  of  $\Gamma$ , for some  $\alpha > 0$ . Since it is no restriction to assume that every region  $R_i$  is contained in  $V_\Gamma^s \cap V_\Gamma^u$ ,  $1 \leq i \leq N$ , see [PT], we will do so.

Let  $\tau: \mathbb{R} \rightarrow W^u(z)$  be a  $C^3$  parametrization, and define a projection  $\pi: \Gamma \rightarrow \bigcup_{i=1}^N R_i \cap I^u$  by taking in each region  $R_i$  the projection along the local stable manifolds into the intersection  $I^u$  with that region,  $1 \leq i \leq N$ . This projection can be extended from  $\Gamma$  to the union of the regions  $R_i$  by projecting along the leaves of the foliation  $\mathcal{F}^s$ . This extension will also be denoted by  $\pi$ . The following result says that for some iterate  $K$ , the map  $F$  can be viewed as expansive along unstable segments.

**Proposition 5.3.** There exist a positive integer  $K$  and a  $C^{1+\alpha}$  map  $\varphi: \bigcup_{i=1}^N \tau^{-1}(I^u \cap R_i) \rightarrow \mathbb{R}$  defined by  $\varphi(x) = \tau^{-1} \circ \pi \circ F^K \circ \tau(x)$  such that  $|\varphi'(x)| > 1$ , for some  $\alpha > 0$ .

*Proof.* For a proof, see Palis and Takens [PT].  $\square$

## 5.2. Proof of the theorem

Let  $J \subset U(R)$  denote an unstable segment. Recall that both end points of  $J$  are on the boundary of the basin boundary transient region  $R$ , and that  $J$  intersects the stable set  $S(R)$ . Recall also that if a point  $x$  in  $R$  eventually leaves  $R$ , then  $x$  is either in  $\text{basin}\{A\}$  or in  $\text{basin}\{B\}$ .

We define for every integer  $k \geq 1$ :

$$C_k(J) = \{x \in J : T_R(x) \geq k\}$$

$$D_k(J) = \{x \in J : T_R(x) = k\}.$$

In particular,  $C_1(J) = J$ . Hence, for each integer  $k \geq 1$  we have  $C_{k+1}(J)$  is the set of points in  $C_k(J)$  whose escape time from  $R$  is at least  $k+1$ ; hence,  $C_{k+1}(J)$  is the set of points in  $J$  that stay in  $R$  under  $F^k$ . The points in  $J$  which stay in  $R$  under all iterates will be denoted by  $C_\infty(J)$ . For every  $k \geq 1$ , we write

$$D_k(J; A) = \{x \in D_k(J) : x \in \text{basin}\{A\}\}$$

$$D_k(J; B) = \{x \in D_k(J) : x \in \text{basin}\{B\}\}.$$

The 'basin boundary property' now implies that for every  $k \geq 1$ :

$$D_k(J) = D_k(J; A) \cup D_k(J; B).$$

Notice that  $\text{basin}\{A\}$  and  $\text{basin}\{B\}$  are disjoint open sets, so that if there are points  $a \in J \cap \text{basin}\{A\}$  and  $b \in J \cap \text{basin}\{B\}$ , then there is a point  $x$  in  $[a, b]$ , with  $T_R(x) = \infty$ . Observe that  $Q$  is a component of  $D_k(J)$  if and only if  $Q$  is a component of either  $D_k(J; A)$  or  $D_k(J; B)$ .

For each  $k \geq 1$  we have:

$$C_k(J) = C_{k+1}(J) \cup D_k(J) = C_{k+1}(J) \cup D_k(J; A) \cup D_k(J; B)$$

$$J = C_{k+1}(J) \cup \bigcup_{j=1}^k D_j(J) = C_{k+1}(J) \cup \bigcup_{j=1}^k D_j(J; A) \cup \bigcup_{j=1}^k D_j(J; B)$$

that is,  $J$  is the union of the set of points  $C_{k+1}(J)$  whose escape time from  $R$  is at least  $k+1$  and, the set of points  $D_j(J)$  whose escape time from  $R$  is  $j$ , and each of those points is either in  $\text{basin}\{A\}$  or in  $\text{basin}\{B\}$ , where  $1 \leq j \leq k$ . We write

$$D_\infty(J) = \bigcup_{k=1}^{\infty} D_k(J) = \bigcup_{k=1}^{\infty} D_k(J; A) \cup \bigcup_{k=1}^{\infty} D_k(J; B).$$

Note that  $C_\infty(J) = \bigcap_{k=0}^{\infty} C_k(J)$ , and  $J = C_\infty(J) \cup D_\infty(J)$ .

Let  $C$  be a component of  $C_k(J)$  that includes a point of  $\text{basin}\{A\}$  (or  $\text{basin}\{B\}$ ). The following result then says that for some fixed positive integer  $s$  (depending only on  $F$  and  $R$ ),  $C$  contains a component of  $\bigcup_{i=0}^{s-1} D_{k+i}(J; A)$  (or  $\bigcup_{i=0}^{s-1} D_{k+i}(J; B)$ ). In particular,  $s$  does not depend on  $k$ . The following lemma (basin boundary combinatorial lemma) is used to prove the 'basin boundary geometric lemma' which follows.

**Basin boundary combinatorial lemma.** Let  $X$  denote either  $A$  or  $B$ . There exists an integer  $s \geq 1$  such that for every unstable segment  $J$  and for each integer  $k \geq 1$  and every component  $C$  of  $C_k(J)$ , the following holds.

If  $C$  includes a point of  $\text{basin}\{X\}$ , then there is an integer  $i$ ,  $k \leq i < k+s$  and a component  $D$  of  $D_i(J; X)$  such that  $D \subset C$ .

*Proof.* Let  $U$  be a neighbourhood of  $\text{Inv}(R)$  on which a  $C^{1+\alpha}$  stable foliation  $\mathcal{F}^s$  exists, for some  $\alpha > 0$ . Select the minimal integer  $v \geq 1$  such that for each basic set  $\Gamma$  of  $F^v$  the following holds, either  $\Gamma$  is a fixed point or  $\Gamma$  is a non-trivial non-periodic basic set. For each non-periodic basic set  $\Gamma$  of  $F^v$ , let  $I_\Gamma^u$  and the regions  $R_i(\Gamma)$ ,  $1 \leq i \leq N(\Gamma)$ , be as in proposition 5.2, and let  $U_\Gamma$  be an open neighbourhood of  $\Gamma$  such that (1)  $\bigcup_{i=1}^{N(\Gamma)} R_i(\Gamma) \subset U_\Gamma \subset U$ , (2) the set  $\tau_\Gamma^{-1}(I_\Gamma^u \cap U_\Gamma)$  consists of  $N(\Gamma)$  open intervals and its closure consists of  $N(\Gamma)$  disjoint intervals, and (3) the map  $\varphi_\Gamma$  in proposition 5.3 may be extended to  $\tau_\Gamma^{-1}(I_\Gamma^u \cap U_\Gamma)$ . For each trivial basic set  $\Gamma$ , let  $U_\Gamma$  be an open neighbourhood of  $\Gamma$  in  $U$  such that  $U_\Gamma$  does not intersect  $U_\Lambda$ , for each basic set  $\Lambda$  in  $\text{Inv}(R) \setminus \Gamma$ .

Let  $L_1^\Gamma, \dots, L_{N(\Gamma)}^\Gamma$  be the components of  $\tau_\Gamma^{-1}(I_\Gamma^\# \cap U_\Gamma)$ ; these finitely many components are open intervals in  $\mathbb{R}$ . Select the minimal integer  $K(\Gamma) \geq 1$  such that the map  $\varphi_\Gamma: \tau_\Gamma^{-1}(I_\Gamma^\# \cap U_\Gamma) \rightarrow \mathbb{R}$  defined by  $\varphi_\Gamma(x) = \tau_\Gamma^{-1} \circ \pi \circ F^{K(\Gamma) \cdot v} \circ \tau_\Gamma(x)$  satisfies  $|\varphi_\Gamma'(x)| > 1$ . Define the map  $\psi_\Gamma: \tau_\Gamma^{-1}(I_\Gamma^\# \cap U_\Gamma) \rightarrow \mathbb{R}$  by  $\psi_\Gamma(x) = \tau_\Gamma^{-1} \circ \pi \circ F^v \circ \tau_\Gamma(x)$ . Now we define the  $N(\Gamma) \times N(\Gamma)$  matrix  $A_\Gamma$  by

$$A_\Gamma(i, j) = \begin{cases} 1 & \text{if } \psi_\Gamma(L_i^\Gamma) \supset L_j^\Gamma \\ 0 & \text{otherwise} \end{cases}$$

for all  $1 \leq i, j \leq N(\Gamma)$ . Since  $\Gamma$  is a non-trivial non-periodic basic set of  $F^v$ , the matrix  $A_\Gamma$  is primitive. Choose the minimal integer  $m(\Gamma) \geq 1$  such that all the entries of the matrix  $A_\Gamma^{m(\Gamma)}$  are positive.

We define the integer  $s(\Gamma)$  as follows. If  $\Gamma$  is a non-trivial non-periodic basic set, then define  $s(\Gamma) = m(\Gamma) \cdot v$ , and if  $\Gamma$  is a fixed point of  $F^v$  define  $s(\Gamma) = v$ . Now, let  $s$  be the smallest common multiple of  $\{s(\Gamma) : \Gamma \text{ is a basic set of } F^v\}$ .

Let  $m(R)$  be the number of basic sets of  $F^v$  in  $\text{Inv}(R)$ , and write  $\text{Inv}(R) = \bigcup_{k=1}^{m(R)} \Gamma_k$ . We associate with  $\text{Inv}(R)$  a directed graph  $G_v$  as follows:  $G_v$  consists of the points  $\Gamma_k$ ,  $1 \leq k \leq m(R)$ , and there exist a path from  $\Gamma_i$  to  $\Gamma_j$  if there exists a point  $z \in \Gamma_i$  such that  $W^u(z) \cap W^s(\Gamma_j)$  is non-empty. Notice that for each  $k$ ,  $1 \leq k \leq m(R)$  there exists a path in  $G_v$  from  $\Gamma_k$  to itself.

Let  $J$  be an arbitrarily chosen unstable segment. Select an integer  $\xi \geq 1$  such that  $C_\xi(J)$  is contained in  $U$ . Let  $\tilde{N}$  denote the number of components of  $C_\xi(J)$ , that is,  $C_\xi(J) = \bigcup_{i=1}^{\tilde{N}} C_{\xi,i}(J)$ . From the definition of the matrices associated with the non-trivial basic sets, the directed graph  $G_v$  associated with  $\text{Inv}(R)$ , and the choice of the integer  $s$ , and using the techniques in [Nu1] and [Nu2], we can associate a  $(0, 1)$ -matrix  $M_J$  with  $C_\xi(J)$ , which is defined by

$$M_J(i, j) = \begin{cases} 1 & \text{if } \pi_J \circ F^v(C_{\xi,i}(J)) \supset C_{\xi,j}(J) \\ 0 & \text{otherwise} \end{cases}$$

for all  $1 \leq i, j \leq \tilde{N}$ , where  $\pi_J$  is the projection on  $J$  along the stable leaves.

We will assume that the  $C_{\xi,i}$ 's are numbered in such a way that the matrix  $M_J$  is written in the normal form, that is,

$$M_J = \begin{bmatrix} M_{11} & 0 & \dots & 0 \\ M_{21} & M_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ M_{m1} & \dots & \dots & M_{mm} \end{bmatrix}$$

where each  $M_{kk}$  is an  $N_k \times N_k$  matrix which is either irreducible (that is, for each pair  $(i, j)$  there exists  $t \in \mathbb{N}$  such that the  $(i, j)$ th entry of the matrix  $(M_{kk})^t$  is positive,  $1 \leq i, j \leq N_k$ ) or a  $1 \times 1$  null matrix,  $1 \leq k \leq m$  and  $\sum_{k=1}^m N_k = \tilde{N}$  for some  $m$ ,  $1 \leq m \leq \tilde{N}$ . This assumption on the  $C_{\xi,i}$ 's is no restriction, since for every non-negative square matrix  $B$  there is a permutation matrix  $P$  such that  $PBP^T$  has the normal form (see Berman and Plemmons [BP]). In particular, each irreducible  $M_{kk}$  is primitive, and if  $N_k \geq 2$  then  $M_{kk}$  equals  $A_\Gamma$  for some non-trivial nonperiodic basic set  $\Gamma$  in  $\text{Inv}(R)$ , and from the choice of the integer  $s$  it follows that all the entries of  $(M_{kk})^s$  are positive.

Let  $X$  denote either  $A$  or  $B$ . Let integer  $k \geq 1$  be given. Let  $C$  be any component  $C_k(J)$ , and assume that  $C$  includes a point of  $\text{basin}(X)$ . We first assume that  $k \geq \xi$ . The definition of  $M_J$ , the choice of  $s$ , and the results in [Nu1] and [Nu2] yields that

there exists an integer  $i$ ,  $k \leq i < k+s$  and a component  $D$  of  $D_i(J; X)$  such that  $D \subset C$ . This result together with the definitions of  $A_k(J)$  and  $D_i(J; X)$  imply immediately that also for  $1 \leq k \leq \xi - 1$  one has that  $C$  includes a component of  $D_i(J; X)$  for some  $i$ ,  $k \leq i < k+s$ .

Since  $J$  was arbitrarily given, we have shown the following. There exists an integer  $s \geq 1$  such that for every unstable segment  $J$  and, for each integer  $k \geq 1$  and every component  $C$  of  $C_k(J)$ , the following holds. If  $C$  includes a point of  $\text{basin}\{X\}$ , then there is an integer  $i$ ,  $k \leq i < k+s$  and a component  $D$  of  $D_i(J; X)$  such that  $D \subset C$ , where  $X$  denotes either  $A$  or  $B$ . This completes the proof of the basin boundary combinatorial lemma.  $\square$

From now on, let  $s$  be as in the 'basin boundary combinatorial lemma', and let  $G = F^s$ . We now consider the escape time of points under  $G$ . For every point  $x$  in  $R$ , the escape time  $T_R^G(x)$  of  $x$  under  $G$  is defined by  $T_R^G(x) = \min\{n \geq \mathbb{N} : G^n(x) \notin R\}$  and  $T_R^G(x) = \infty$  if  $G^n(x) \in R$  for all  $n \geq 1$ . We say that  $T_R^G(x) = 0$  if  $x \notin R$ .

We define for every integer  $k \geq 1$ :

$$C_k^G(J) = \{x \in J : T_R^G(x) \geq k\}$$

$$D_k^G(J; A) = \{x \in J : T_R^G(x) = k \text{ and } x \in \text{basin}\{A\}\}$$

$$D_k^G(J; B) = \{x \in J : T_R^G(x) = k \text{ and } x \in \text{basin}\{B\}\}.$$

Hence, for each integer  $k \geq 1$  we have  $C_{k+1}^G(J)$  is the set of points in  $C_k^G(J)$  whose escape time under  $G$  from  $R$  is at least  $k+1$ ; hence,  $C_{k+1}^G(J)$  is the set of points in  $J$  that stay in  $R$  under  $G^k$ . The points in  $J$  which will stay in  $R$  under all iterates will be denoted by  $C_\infty^G(J)$ . For each  $k \geq 1$  we have:

$$C_k^G(J) = C_{k+1}^G(J) \cup D_k^G(J; A) \cup D_k^G(J; B)$$

$$J = C_{k+1}^G(J) \cup \bigcup_{j=1}^k D_j^G(J; A) \cup \bigcup_{j=1}^k D_j^G(J; B)$$

that is,  $J$  is the union of the set of points  $C_{k+1}^G(J)$  whose escape time under  $G$  from  $R$  is at least  $k+1$  and, the set of points  $D_j^G(J; A)$  in  $\text{basin}\{A\}$  (respectively,  $D_j^G(J; B)$  in  $\text{basin}\{B\}$ ) whose escape time under  $G$  from  $R$  is  $j$ , where  $1 \leq j \leq k$ . We write

$$D_\infty^G(J) = \bigcup_{k=1}^{\infty} D_k^G(J; A) \cup \bigcup_{k=1}^{\infty} D_k^G(J; B)$$

Note that  $C_\infty^G(J) = \bigcap_{k=0}^{\infty} C_k^G(J)$ , and  $J = C_\infty^G(J) \cup D_\infty^G(J)$ .

**Lemma 5.4.** For every integer  $k \geq 1$ , we have:

- (1)  $D_k^G(J; A) = \bigcup_{i=k-s+1}^{ks} D_i(J; A)$ ;  $D_k^G(J; B) = \bigcup_{i=k-s+1}^{ks} D_i(J; B)$ ;
- (2)  $C_\infty^G(J) = C_\infty(J)$  and  $D_\infty^G(J) = D_\infty(J)$ ;
- (3) each component of  $D_\infty^G(J)$  belongs to either  $\text{basin}\{A\}$  or  $\text{basin}\{B\}$ .

*Proof.* The proof is left to the reader.  $\square$

Note that the set  $D_\infty^G(J)$  is the set of points  $x \in J$  with finite escape time (that is,  $T_R^G(x) < \infty$ ). The following result says that, if the value of the escape time map  $T_R^G$  changes then it changes in steps of 1. Denote the length of a connected subset  $L \subset J$  by  $\rho(L)$ .

*T-jump property.* For every  $x \in J$  with  $T_R^G(x) < \infty$ , there exists  $\varepsilon > 0$  such that each  $y \in J$  with  $\rho([x, y]_J) < \varepsilon$  satisfies  $|T_R^G(x) - T_R^G(y)| \leq 1$ .

*Proof.* Apply lemma 5.4 and the *T-jump lemma* in [NY2].  $\square$

The following lemma for  $G$  implies that if an unstable segment  $\gamma$  has a component  $C$  of  $C_k^G(\gamma)$  that intersects  $\text{basin}\{X\}$ , then there is a point  $p$  of  $C_k^G(\gamma) \cap \text{basin}\{X\}$  with escape time  $k$ , and the length of the component  $D$  of  $D_k^G(\gamma; X)$  including  $p$  is at least  $\delta \cdot \rho(C)$ .

*Basin boundary geometric lemma.* Let  $X$  denote either  $A$  or  $B$ . There exists  $\delta > 0$ , such that for every unstable segment  $J$ , and for each integer  $k \geq 1$  and every component  $C$  of  $C_k^G(J)$ , we have:

If  $C$  includes a point of  $\text{basin}\{X\}$ , then there is a component  $D$  of  $D_k^G(J; X)$  such that  $D \subset C$  and  $\rho(D)/\rho(C) \geq \delta$ .

*Proof.* From the geometric lemma II in [NY2] applied to  $G$ , there exists  $\delta > 0$  such that for every  $J$  in  $U(R)$ , and for every integer  $k \geq 1$ , the following holds:

- (1) each component of  $C_k^G(J)$  contains components of  $C_k^G(J)$  and  $C_{k+1}^G(J)$ ; and
- (2) if  $C$  is any component of  $C_k^G(J)$ , then every component  $D$  of  $D_k^G(J) \cap C$  satisfies  $\rho(D)/\rho(C) \geq \delta$ , and every component  $U$  of  $C_{k+1}^G(J) \cap C$  satisfies  $\rho(U)/\rho(C) \geq \delta$ .

Let  $X$  denote either  $A$  or  $B$ , and let  $J$  be any unstable segment. Let integer  $k \geq 1$  and component  $C$  of  $C_k^G(J)$  be given. Assume that  $C$  includes a point of  $\text{basin}\{X\}$ .

Applying the basin boundary combinatorial lemma yields that there exists a component  $D$  of  $D_k^G(J; X)$  such that  $D \subset C$ . From the geometric lemma II in [NY2], since  $D$  is a component of  $D_k^G(J)$ , and the definition of  $\delta$ , we obtain  $\rho(D)/\rho(C) \geq \delta$ . Since  $J$ ,  $k$  and  $C$  are assumed to be given arbitrarily, we conclude for each unstable segment  $J$ , for each integer  $k \geq 1$  and every component  $C$  of  $C_k^G(J)$ , if  $C$  includes a point of  $\text{basin}\{X\}$ , then there is a component  $D$  of  $D_k^G(J; X)$  such that  $D \subset C$  and  $\rho(D)/\rho(C) \geq \delta$ . This completes the proof of the basin boundary geometric lemma.  $\square$

From now on, we fix  $\delta$  as in the basin boundary geometric lemma. Before we prove the theorem, we present a non-intertwining property for the escape time map as well as an auxiliary observability result for accessible straddle pair sequences. We call a pair  $\{p, q\}$  a *balanced pair* if  $T_R^G(p) = T_R^G(q)$ .

*Non-intertwining lemma.* Let  $\{p, q\}$  be a balanced pair, let  $P$  be a  $\delta^2$ -refinement of  $\{p, q\}$ , and assume that  $T_R^G(x_i) \geq T_R^G(p)$  for every  $x_i$  in  $P$ . If each point of  $P$  is in  $\text{basin}\{A\}$  then  $[p, q]_J$  is contained in  $\text{basin}\{A\}$ .

*Proof.* Let  $\{p, q\}$  and  $P$  be as in the lemma. Assume that each point of  $P$  is in  $\text{basin}\{A\}$ . Write  $m = \min\{T_R^G(x) : x \in [p, q]_J\}$ . The assumptions ' $T_R^G(x_i) \geq T_R^G(p)$  for all  $x_i \in P$ ', ' $P$  is a  $\delta^2$ -refinement of  $\{p, q\}$ ', together with basin boundary geometric lemma yields that  $m = T_R^G(p)$ . Hence,  $[p, q]_J$  is contained in a component of  $C_m^G(J)$ . If there exists a component  $D$  of  $D_m^G(J)$  including  $[p, q]_J$ , then  $D$  is a component of  $D_m^G(J; A)$ , and we are done. Therefore, from now on, we assume that  $[p, q]_J$  is not contained in a component of  $D_m^G(J)$ . This implies that there are at least one

component of  $C_{m+1}^G(J)$  in the interior of  $[p, q]_J$ , and at least two components of  $D_m^G(J)$  which have a non-empty intersection with  $[p, q]_J$ .

Let  $D$  be a component of  $D_m^G(J)$  such that  $D \cap [p, q]_J$  is non-empty. By the basin boundary geometric lemma, we have  $\rho(D)/\rho([p, q]_J) > \delta$ . Since  $P$  is an  $\varepsilon$ -refinement of  $\{p, q\}$ , it follows that  $P \cap D$  is non-empty. This fact and the assumption that each point of  $P$  is in  $\text{basin}\{A\}$  imply that  $D$  is a component of  $D_m^G(J; A)$ . This implies that  $[p, q]_J \cap D_m^G(J)$  is contained in  $\text{basin}\{A\}$ .

Let  $C$  be any component of  $C_{m+1}^G(J)$  in the interior of  $[p, q]_J$ . Applying the basin boundary geometric lemma we get that  $\rho(C)/\rho([p, q]_J) > \delta$ . If  $C$  includes a point of  $\text{basin}\{B\}$ , then  $C$  includes a component  $D$  of  $D_{m+1}^G(J; B)$ , and by the basin boundary geometric lemma, we have

$$\rho(D)/\rho([p, q]_J) = (\rho(D)/\rho(C)) \cdot (\rho(C)/\rho([p, q]_J)) > \delta^2.$$

Hence, if  $C$  includes a point of  $\text{basin}\{B\}$  then every  $\delta^2$ -refinement of  $\{p, q\}$  includes a point of  $\text{basin}\{B\}$ . Since  $P$  is an  $\delta^2$ -refinement of  $\{p, q\}$  and  $P$  does not contain a point of  $\text{basin}\{B\}$ , it follows that  $C$  includes no point of  $\text{basin}\{B\}$ . Since  $C$  is arbitrary, we get that each component of  $C_{m+1}^G(J)$  that is in  $[p, q]_J$  contains no point of  $\text{basin}\{B\}$ . Therefore  $[p, q]_J \cap C_{m+1}^G(J)$  is contained in  $\text{basin}\{A\}$ .

Because of  $[p, q]_J = ([p, q]_J \cap D_m^G(J)) \cup ([p, q]_J \cap C_{m+1}^G(J))$  the conclusion is that  $[p, q]_J$  is contained in  $\text{basin}\{A\}$ . This completes the proof of the non-intertwining lemma.  $\square$

*Basin boundary observability lemma.* Let  $P$  be a  $\delta^3/3$ -refinement of a straddle pair  $\{a_0, b_0\}$ , and assume  $T_R^G(x_i) \geq T_R^G(a_0)$  for every  $x_i$  in  $P$ . Let  $\{a_0, b_1\}$  be the straddle pair in  $P$ , in which  $b_1$  is selected as in the accessible basin boundary refinement procedure. Let  $a_1^\dagger$  be defined as in the improved version of the accessible basin boundary refinement procedure. If  $P$  is a  $\delta^3$ -refinement of  $\{a_0, b_1\}$ , then  $[a_0, a_1^\dagger]_J$  is in  $\text{basin}\{A\}$ , and  $T_R^G(a_1^\dagger) = T_R^G(a_0) + 1$ .

*Proof.* Let  $P$ ,  $\{a_0, b_1\}$ , and  $a_1^\dagger$  be as in the lemma, and assume that  $P \cap [a_0, b_1]_J$  is an  $\varepsilon$ -refinement of  $\{a_0, b_1\}$ , where  $\varepsilon = \delta^3$ . Let  $m = \min\{T_R^G(x) : x \in [a_0, b_1]_J\}$ . Let  $a_1^0$  and  $a_1^+$  be defined as in the improved version of the accessible basin boundary refinement procedure.

The assumptions ' $T_R^G(x_i) \geq T_R^G(a_0)$  for all  $x_i \in P$ ', ' $P \cap [a_0, b_1]_J$  is an  $\varepsilon$ -refinement of  $\{a_0, b_1\}$ ', together with the basin boundary geometric lemma yields  $m = T_R^G(a_0)$ . Hence,  $[a_0, b_1]_J$  is contained in a component of  $C_m^G(J)$ .

By definition, we have  $a_0 \leq a_1^0$ . We show first that  $[a_1^0, a_1^\dagger]_J$  is contained in  $\text{basin}\{A\}$ . Applying the  $T$ -jump property and the basin boundary geometric lemma we obtain that there exists a component  $D$  of  $C_{m+1}^G(J; A)$  such that  $D$  is in the interior of  $[a_1^0, b_1]_J$ , and  $\rho(D)/\rho([a_0, b_1]_J) > \delta$ . Therefore,  $a_1^+$  exists and  $T_R^G(a_1^+) = m + 1$ . The definition of  $a_1^+$  and lemma 5.4 imply that  $[a_1^0, a_1^+]_J$  is contained in  $\text{basin}\{A\}$ . Recall that  $\{a_1^+, a_1^\dagger\}$  is a balanced pair, that is,  $T_R^G(a_1^\dagger) = T_R^G(a_1^+)$ . If  $a_1^+$  and  $a_1^\dagger$  are in the same component of  $D_{m+1}^G(J)$  then  $[a_1^+, a_1^\dagger]_J$  is in  $\text{basin}\{A\}$ , and we get that  $[a_1^0, a_1^\dagger]_J$  is in  $\text{basin}\{A\}$ . Now assume that  $a_1^+$  and  $a_1^\dagger$  are in different components of  $D_{m+1}^G(J)$ . Then,  $[a_1^+, a_1^\dagger]_J$  includes at least one component  $C$  of  $C_{m+2}^G(J)$  in its interior, and by the basin boundary geometric lemma we have  $\rho(C)/\rho([a_1^+, a_1^\dagger]_J) > \delta$ . This implies that  $P \cap [a_1^+, a_1^\dagger]_J$  is a  $\delta^2$ -refinement of  $\{a_1^+, a_1^\dagger\}$ . Applying the non-intertwining lemma yields  $[a_1^+, a_1^\dagger]_J$  is in  $\text{basin}\{A\}$ , and we obtain also in this case that  $[a_1^0, a_1^\dagger]_J$  is contained in  $\text{basin}\{A\}$ . We conclude:  $\text{basin}\{A\}$  includes  $[a_1^0, a_1^\dagger]_J$ , and  $T_R^G(a_1^\dagger) = T_R^G(a_0) + 1$ .

If  $a_0 = a_1^0$ , then it follows immediately from the conclusion above that  $[a_0, a_1^0]_J$  is contained in  $\text{basin}\{A\}$ . From now on, we assume  $a_0 < a_1^0$ . Recall that  $\{a_0, a_1^0\}$  is a balanced pair. If  $a_0$  and  $a_1^0$  are in the same component of  $D_m^G(J)$ , then  $[a_0, a_1^0]_J$  is in  $\text{basin}\{A\}$ . If  $a_0$  and  $a_1^0$  are in different components of  $D_m^G(J)$ , then  $[a_0, a_1^0]_J$  includes at least one component  $C$  of  $C_{m+1}^G(J)$ . Since  $\rho([a_0, a_1^0]_J)/\rho([a_0, b_1]_J) > \rho(C)/\rho([a_0, a_1^0]_J) > \delta$  and  $P \cap [a_0, a_1^0]_J$  is a  $\delta^2$ -refinement of  $\{a_0, a_1^0\}$ , applying the non-intertwining lemma we obtain  $[a_0, a_1^0]_J$  is in  $\text{basin}\{A\}$ . Since  $[a_1^0, a_1^0]_J$  is in  $\text{basin}\{A\}$ , the conclusion is that  $[a_0, a_1]_J$  is contained in  $\text{basin}\{A\}$ . This completes the proof of the basin boundary observability lemma.  $\square$

*Proof of the theorem.* Let  $\delta$  be as in the basin boundary geometric lemma, and choose  $\varepsilon = \delta^3$ . Let  $\{[a_n, b_n]_J\}_{n \geq 0}$  be an accessible straddle segment sequence, that is,  $\{a_0, b_0\}$  is a straddle pair and  $\{a_n, b_n\}$  is obtained by the improved version of the accessible basin boundary refinement procedure for all  $n \geq 1$ . For  $n \geq 0$ , let  $P_n$  be an  $\varepsilon/3$ -refinement of  $\{a_n, b_n\}$ , and let  $m_n$  be as in the improved version of the accessible basin boundary refinement procedure. By the basin boundary geometric lemma we obtain  $m_n = \min\{T_R^G(x) : x \in [a_n, b_{n+1}]_J\}$ .

We will show that there exists an integer  $N \geq 0$  such that for every integer  $n \geq N$  the following properties hold. (P1)  $T_R^G(a_n) = m_n$ , (P2)  $|T_R^G(a_{n+1}) - T_R^G(a_n)| \leq 1$ , and (P3)  $[a_n, a_{n+1}]_J$  is contained in  $\text{basin}\{A\}$ . Notice that we do not claim that  $|T_R^G(x) - T_R^G(a_n)| \leq 1$  for all  $x \in [a_n, a_{n+1}]_J$ , where  $n \geq N$ .

From the  $T$ -jump property and the basin boundary geometric lemma, together with the assumption that  $\{[a_n, b_n]_J\}_{n \geq 1}$  is obtained using the accessible basin boundary procedure, we have that if  $T_R^G(a_n) > m_n$  then  $T_R^G(a_{n+1}) = m_n$ , for each  $n \geq 0$ . This property implies that there exists a minimal integer  $N \geq 0$  such that  $T_R^G(x_i) \geq m_N = T_R^G(a_N)$  for each  $x_i \in P_N$ . Hence, (P1) holds for  $N$ . We now show that (P2) and (P3) hold for this integer  $N$ .

*Case 1.*  $P_N$  is not an  $\varepsilon$ -refinement of  $\{a_N, b_{N+1}\}$ . Then  $a_{N+1} = a_N$ , so  $[a_N, a_{N+1}]_J$  is contained in  $\text{basin}\{A\}$  and  $T_R^G(x_i) \geq m_{N+1} = T_R^G(a_{N+1})$  for each  $x_i \in P_{N+1}$ . Therefore, (P3) holds, while (P2) is obvious since  $a_N = a_{N+1}$ .

*Case 2.*  $P_N$  is an  $\varepsilon$ -refinement of  $\{a_N, b_{N+1}\}$ . The basin boundary observability lemma implies (P3) since  $[a_N, a_{N+1}]_J$  is contained in  $\text{basin}\{A\}$ . It also implies (P2) since  $T_R^G(x_i) \geq m_{N+1} = T_R^G(a_{N+1}) = T_R^G(a_N) + 1$  for each  $x_i \in P_{N+1}$ .

By induction, one obtains the desired result. This completes the proof of the theorem.

## 6. The numerical procedure and related numerical methods

### 6.1. The dynamic problem

Now we return to the 'dynamic' problem stated in the introduction, namely, to describe a procedure for finding a numerical trajectory on the basin boundary which is accessible from  $\text{basin}\{A\}$ . (Recall that the basin boundary of  $\text{basin}\{A\}$  is the boundary of the closure of  $\text{basin}\{A\}$ .) We assume we are given a straight line segment that intersects the basin boundary transversally and has one end point in

basin $\{A\}$  and the other end point in basin $\{B\}$ . In the statement of the results, we assume that a straddle pair and its  $\varepsilon$ -refinement lie in a connected subset of an unstable segment, and that all unstable segments intersect the basin boundary transversally. However, from our proof of the theorem it follows that a similar result holds if we replace the unstable segment by a straight line segment so that we assume that every  $\varepsilon$ -refinement of a straddle pair  $\{a, b\}$  is in the straight line segment  $[a, b]$  from  $a$  to  $b$ , and that  $[a, b]$  intersects the basin boundary transversally.

A straight line segment  $[a, b]$  *straddles* the stable manifold of a point  $P$  if  $[a, b]$  intersects this manifold transversally. In the cases we study, that is,  $a \in \text{basin}\{A\}$  and  $b \in \text{basin}\{B\}$ , the stable manifold of  $P$  will be replaced by a (fractal) basin boundary and  $[a, b]$  will straddle a subset of the basin boundary. Furthermore, in practice  $[a, b]$  will be very short and will be extremely close to the invariant set  $\text{Inv}(R)$ .

The numerical procedure goes as follows.

(1) Choose (with some experimenting) a straddle pair  $\{a, b\}$  and let  $I$  denote the line segment from  $a$  to  $b$ .

(2) Apply the accessible basin boundary refinement procedure (that is, refine and choose a new straddle pair  $\{x, y\}$  in  $I$  and then replace  $I$  by the straight line segment from  $x$  to  $y$ . Repeat this process until the length of  $I$  is less than some distance  $\sigma$  (for example,  $\sigma = 10^{-8}$ ). If the initial  $a$  and  $b$  are less than  $\sigma$  apart, then the pair is not changed.

Given any initial straddle pair  $\{a, b\}$ , we will write  $\{a_0, b_0\} = \text{ABS}_\sigma(\{a, b\})$ , for the straddle pair resulting from step 2. Note that  $\|a_0 - b_0\| < \sigma$ . 'ABS' is an abbreviation of 'accessible basin boundary straddle refinement'.

(3) For each integer  $n \geq 0$ , and straddle pair  $\{a_n, b_n\}$  such that  $\|a_n - b_n\| < \sigma$ , compute the refinement for the image pair  $\{F(a_n), F(b_n)\}$ , and write

$$\{a_{n+1}, b_{n+1}\} = \text{ABS}_\sigma(F(a_n), F(b_n)).$$

Thus we obtain a sequence  $\{\{a_n, b_n\}\}_{n \geq 0}$  of straddle pairs. Note that only  $F(a_n)$  and  $F(b_n)$  and  $\sigma$  are relevant to the computation of  $\{a_{n+1}, b_{n+1}\} = \text{ABS}_\sigma(\{F(a_n), F(b_n)\})$ , since  $\text{ABS}_\sigma(\{F(a_n), F(b_n)\})$  is a straddle pair in the line segment from  $F(a_n)$  to  $F(b_n)$ .

Write  $I_n$  for the line segment from  $a_n$  to  $b_n$ . Since the system is hyperbolic and the matrix of the second partial derivatives  $D^2F$  will be bounded on the closure of the region  $R$ , there will be a bound on the curvature of the curve  $F(I_n)$ , and  $F(I_n)$  will deviate from the straight line segment  $L_n$  from  $F(a_n)$  to  $F(b_n)$  by an amount proportional to  $|L_n|^2$ , where  $|L_n|$  denotes the length of  $L_n$ .

We thus obtain a trajectory of tiny straight line segments  $I_n$  and to the precision of the computer (about  $10^{-14}$ ) we usually have  $I_{n+1} \subset F(I_n)$ , and selecting any point  $x_n$  from  $I_n$ , perhaps the midpoint, we have that  $|x_{n+1} - F(x_n)|$  is small, typically of the order of  $\sigma$ . Since computers can never be expected to produce true trajectories (except in trivial cases such as fixed points), we may say  $\{x_n\}_{n \geq 0}$  is a numerical trajectory with precision  $\sigma$ . Despite the complexity of the construction, we will refer to  $x_{n+1}$  as the 'iterate' of  $x_n$ . We call the sequence of intervals  $\{I_n\}_{n \geq 0}$  an *accessible basin boundary straddle trajectory* or *ABST trajectory*, and we call the numerical procedure above that generates the sequence  $\{I_n\}_{n \geq 0}$ , the *accessible basin boundary straddle method* or *ABST method*. Notice that each interval straddles a piece of the

basin boundary. After a few iterates, the sequence  $\{x_n\}_{n \geq 0}$  resembles a subset of the non-wandering points in  $R$  which are accessible from  $\text{basin}(A)$ .

In this paper we have shown that our procedure (the accessible basin boundary refinement procedure) is valid in ideal situations. We find that the accessible basin boundary straddle method works well in practice even in less than ideal cases, in particular cases where hyperbolicity seems to fail. If  $\varepsilon$  is chosen too large, then  $\{(a_n, b_n)\}_{n \geq 0}$  would still be a sequence of straddle pairs with  $a_n \in \text{basin}(A)$  and  $b_n \in \text{basin}(B)$ , but the sequence would not be accessible and probably would not settle down to a periodic orbit.

In practice we find that, in most cases we study, the method appears to work well for  $\varepsilon = 1/30$ . In computing the sequence of straddle pairs  $\{a_n, b_n\}$  defined by the accessible basin boundary refinement procedure, once case (2c) holds, then it can be shown that every  $\varepsilon$ -refinement of the proper straddle pair  $\{a, b\}$  includes a proper straddle pair. For the examples in this paper we find that the accessible basin boundary straddle method leads (in all cases but one) to accessible fixed points or periodic points, in agreement with the fact that all the accessible points for two-dimensional saddle-hyperbolic systems are on the stable manifolds of finitely many periodic points. The exceptional case is the example of the complex quadratic map of which the basin boundary is two-dimensionally unstable, and the result due to Newhouse and Palis does not apply in this particular case.

## 6.2. The accessible set on the basin boundary

We have seen above that in many interesting cases our numerical method (accessible basin boundary straddle method) produces a periodic trajectory on the basin boundary that is in  $\text{Inv}(R)$ . If  $P$  is a periodic trajectory in  $\text{Inv}(R)$  that is accessible from  $\text{basin}(A)$ , then all the points on the stable manifold of  $P$  are accessible from  $\text{basin}(A)$ . Therefore, we need a numerical method that produces the stable manifold of a periodic point. In [YKY] a procedure has been presented that can be used for the calculation of stable manifolds of saddle periodic points of the diffeomorphism  $F$ . The calculation can be made with a guaranteed accuracy, in particular, it can be used to calculate the pieces of the stable manifolds of the periodic points that we find. As illustration, we present in figure 9 the stable

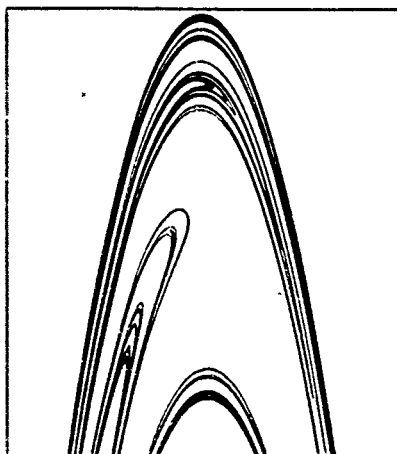


Figure 9. The stable manifold of the fixed point of the Hénon map (with  $\rho = 1.405$ ,  $\mu = -0.3$ ) that is accessible from  $\text{basin}(B)$ .

manifold of the period 1 saddle in the example of the Hénon map for which the attractor infinity (attractor  $A$ ) and a period 2 attractor coexist. This stable manifold of the saddle fixed point constitute the accessible set (accessible from  $\text{basin}(A)$ ) on the basin boundary.

### 6.3. Related straddle trajectories

In this subsection we review briefly 'straddle trajectories' that are obtained by methods which are based on refinement procedures such as the bisection procedure [BGOYY], [GNOY], the PIM triple refinement procedure [NY1], [NY2] and the accessible PIM triple refinement procedure [NY2]. The methods were used in the applications presented in section 3 and the refinement procedures above are related to the accessible basin boundary refinement procedure. These straddle methods are numerical methods for obtaining trajectories on the basin boundary and on chaotic saddles. For clarity of the exposition and in order that this paper is self-contained, we describe these methods; see the references above for details.

Straddle methods involve a refinement procedure in which 2 points on a curve segment are replaced by two new points. In some cases the points have different roles. Usually each of the refinement procedures takes a pair of points and returns a pair of points; such a returned pair is on the line segment joining the two points of the original pair. The distance between the two points in the returned pair is smaller than the distance between the points of the original pair. Straddle methods consist of applying the refinement procedure repeatedly until the points in the resulting pair are less than some specified distance  $\sigma$  apart, say  $\sigma = 10^{-8}$ . If the points in the original pair are already less than  $\sigma$  apart, then no refinement is carried out. Next apply the dynamics; that is, apply the map  $F$  to each of the two points of the resulting pair, giving a new pair.

The basic numerical method takes a pair  $\{a_n, b_n\}$  which is separated by at most a distance  $\sigma$ , and applies the map  $F$  to each of the points of this pair. If the new pair  $\{F(a_n), F(b_n)\}$  is separated by less than  $\sigma$ , then it is denoted  $\{a_{n+1}, b_{n+1}\}$ , and otherwise the refinement procedure is applied repeatedly until a pair with separation at most  $\sigma$  is obtained, and it is called  $\{a_{n+1}, b_{n+1}\}$ . However, in order to produce the first pair  $\{a_0, b_0\}$ , the method starts by applying the refinement procedure on the given pair  $\{a, b\}$ , whose points are presumably more than  $\sigma$  apart. Writing  $I_n$  or  $[a_n, b_n]$  for the line segment from  $a_n$  to  $b_n$ , and to the precision of the computer we usually have  $I_{n+1} \subset F(I_n)$ . We call the sequence of tiny straight line segments  $\{I_n\}_{n \geq 0}$  a *straddle trajectory*.

**BST method.** The 'basin boundary dynamic problem' is to develop a numerical method for finding a trajectory on the basin boundary.

The refinement procedure for straddle pairs is particularly simple. Let  $\{\alpha, \beta\}$  be a straddle pair such that  $\alpha \in \text{basin}(A)$  and  $\beta \in \text{basin}(B)$ . We define  $\gamma$  to be the midpoint of the straight line segment  $[\alpha, \beta]$ , that is,  $\gamma = (\alpha + \beta)/2$ . If  $\gamma \in \text{basin}(A)$  then we choose  $\alpha^* = \gamma$ ,  $\beta^* = \beta$ ; otherwise, if  $\gamma \in \text{basin}(B)$  then we choose  $\alpha^* = \alpha$ ,  $\beta^* = \gamma$ . This refinement procedure is also called the *bisection procedure*.

The solution to the 'basin boundary dynamic problem' is the straddle trajectory using the bisection procedure. We call the sequence of tiny straight line segments  $\{I_n\}_{n \geq 0}$  a *basin boundary straddle trajectory* or *BST trajectory*, and we call the

straddle method above that generates the BST trajectory  $\{I_n\}_{n \geq 0}$ , the *basin boundary straddle trajectory method* or *BST method*. Notice that each tiny line segment in a BST trajectory straddles the basin boundary. A BST trajectory typically resembles (after a few iterates) a basic set in the basin boundary.

*SST method.* The 'saddle dynamic restraint problem' is to describe a numerical method for finding a trajectory that remains in a specified transient region for an arbitrarily long period of time.

First, we describe the refinement procedure that is involved in the current straddle method. Let  $\{a, b\}$  be a pair such that  $[a, b]$  intersects  $S(R)$  transversally. The notation  $(x, y, z)$  for a triple means that  $x, y$ , and  $z$  lie on  $[a, b]$  and  $y$  is between  $x$  and  $z$ , and we assume for convenience that the ordering on  $[a, b]$  is such that  $x < y < z$ . For each  $\varepsilon > 0$ , an  $\varepsilon$ -refinement of a triple  $(x, z, y)$  is an  $\varepsilon$ -refinement of  $(x, y)$  such that it includes  $z$ . Let  $(\alpha, \gamma, \beta)$  be a triple on  $[a, b]$ . We call  $(\alpha, \gamma, \beta)$  an *Interior or Maximum triple* if both  $T_R(\gamma) > T_R(\alpha)$  and  $T_R(\gamma) > T_R(\beta)$ ; and we call  $(\alpha, \gamma, \beta)$  a *PIM triple* if  $(\alpha, \gamma, \beta)$  is an Interior Maximum triple and  $\|\beta - \alpha\| < \|b - a\|$ .

Let  $(\alpha, \gamma, \beta)$  be an Interior Maximum triple, and let  $P$  be an  $\varepsilon$ -refinement of  $(\alpha, \gamma, \beta)$ . The procedure that selects in the refinement  $P$  any PIM triple  $(\alpha^*, \gamma^*, \beta^*)$  is called a *PIM triple (refinement) procedure*.

The solution to the 'saddle dynamic restraint problem' is the straddle trajectory using the PIM triple procedure. We call the sequence of tiny straight line segments  $\{I_n\}_{n \geq 0}$  a *saddle straddle trajectory* or *SST trajectory*, and we call the straddle method that generates the SST trajectory  $\{I_n\}_{n \geq 0}$ , the *saddle straddle trajectory method* or *SST method*. Notice that each tiny line segment in an SST trajectory straddles a piece of a (chaotic) saddle. An SST trajectory typically resembles (after a few iterates) a basic set in the chaotic saddle.

*ASST method.* The 'accessible saddle dynamic restraint problem' is to describe a numerical method for finding a trajectory on the stable set  $S(R)$  that is accessible from the transient set  $R \setminus S(R)$ .

The refinement procedure that is involved in the current straddle method is a PIM triple (refinement) procedure in which a PIM triple  $(\alpha^*, \gamma^*, \beta^*)$  is selected from the  $\varepsilon$ -refinement  $P$  of the interior maximum triple  $(a, c, b)$  such that  $[a, \alpha^*]$  is in the transient set  $R \setminus S(R)$  (hence,  $[a, \alpha^*]$  does not intersect the stable set  $S(R)$ ). This refinement procedure is called the *accessible PIM triple (refinement) procedure*.

The solution to the 'accessible saddle dynamic restraint problem' is the straddle trajectory using the accessible PIM triple procedure. We call the straddle trajectory  $\{I_n\}_{n \geq 0}$  an *accessible saddle straddle trajectory* or *ASST trajectory*, and we call the straddle method that generates the ASST trajectory  $\{I_n\}_{n \geq 0}$ , the *accessible saddle straddle trajectory method* or *ASST method*. An ASST trajectory typically resembles (after a few iterates) a subset of the non-wandering points in  $R$  which are accessible from the transient set  $R \setminus S(R)$ .

In most cases that we have investigated we find that every  $\varepsilon$ -refinement of two points  $\{a, b\}$ , when  $\varepsilon$  is chosen to be  $1/30$ , includes several PIM triples. In [NY1], [NY2] we find that the ASST method leads to accessible fixed points or periodic points, which is in agreement with the fact that all the accessible points for two dimensional hyperbolic systems are on the stable manifolds of finitely many periodic points. In [NY2] we have shown that the two PIM triple procedures are valid in ideal situations (hyperbolic systems). We find SST and ASST methods work well in

practice even in less than ideal cases. From the examples in [NY1], we have seen that the SST method works quite well for a variety of dynamical systems.

Most pictures in section 3 for which one of the numerical straddle procedures has been applied in order to obtain a single numerical trajectory, have been obtained by selecting  $\varepsilon = 1/30$  as default value, and neglecting the first 10 iterates. We chose  $\varepsilon$  to be somewhat smaller (0.01) in the ABST method for the Hénon map (parameter values  $\rho = 2.66$ ,  $\mu = 0.3$ ).

#### 6.4. Shadowing

It is important to ask if such straddle trajectories obtained by one of the straddle methods (BST method, SST method, ASST method, or ABST method) represent true trajectories of the system. In other words, does there exist a true trajectory of the system that shadows (i.e. stays close to) the numerical trajectory obtained by a straddle method? When a map is sufficiently hyperbolic on the invariant set in question, Bowen [B] obtained a result saying that each noisy trajectory in the non-wandering set can be shadowed by a true trajectory if the perturbation is small; see [B] for the precise statement. Recall that  $\text{Inv}(R)$  satisfies the 'no cycle condition' if whenever basic sets  $\Gamma_{k(1)}, \dots, \Gamma_{k(M)}$  is a sequence of basic sets in  $\text{Inv}(R)$  for which the stable set of  $\Gamma_{k(i)}$  has a non-empty intersection with the unstable set of  $\Gamma_{k(i+1)}$  for all  $1 \leq i < k(M)$ , then the stable set of  $\Gamma_{k(M)}$  does not intersect the unstable set of  $\Gamma_{k(1)}$ . Assuming  $\text{Inv}(R)$  satisfies the 'no cycle condition' and  $\delta$  is sufficiently small, we can show that every BST or SST trajectory of a two dimensional uniformly hyperbolic system with a fractal basin boundary or a chaotic saddle, obtained by the BST method and SST method respectively, can be shadowed by a true trajectory (for as long as the saddle straddle trajectory can be computed).

### 7. Concluding remarks

#### 7.1. Higher-dimensional systems

One of the ingredients in the analysis of the validity of the accessible basin boundary procedure in dimension two, is the existence of a  $C^{1+\alpha}$  foliation  $\tilde{\mathcal{F}}$  on a neighbourhood of a basic set. The proofs of the basin boundary geometric lemma and the basin boundary combinatorial lemma require the existence of such a stable foliation (see also the proof of geometric lemma II in [NY2], on which the proof of the basin boundary geometric lemma is heavily based). For  $d = 2$ , the existence of such a foliation is guaranteed by a result due to De Melo [M]. Unfortunately, the existence of a foliation  $\tilde{\mathcal{F}}$  on a neighbourhood of a basic set in higher dimensions is not known, see e.g. [PT].

Let from now on, the dimension  $d \geq 3$ . Let  $F$  be an Axiom A diffeomorphism, let  $R$  be a basin boundary region such that  $\dim E^u = 1$ , and assume that for each basic set  $\Gamma$  in  $\text{Inv}(R)$  there exists a  $C^{1+\alpha}$  stable foliation  $\tilde{\mathcal{F}}$  on a neighbourhood of  $\Gamma$ , for some  $\alpha > 0$ . Then the conclusion of the theorem is again valid. The proof is almost the same; instead of proposition 5.2 one should use the properties of Markov partitions of basic sets; see Bowen [B].

### 7.2. Order of differentiability of the diffeomorphism

We assumed that the diffeomorphism  $F$  is  $C^3$ . This assumption implied the existence of a  $C^{1+\alpha}$  expanding map, for some  $\alpha > 0$ , in proposition 5.3. If  $F$  is of class  $C^2$ , then it is known that such an expanding map is  $C^1$ . We would like to point out, that the Hölder exponent  $\alpha$  is only used to obtain (2) in the proof of the Geometric lemma I in [NY2]; the proof of the basin boundary Geometric lemma depends indirectly on this result. Fortunately, we can prove Geometric lemma I in [NY2] (in particular the property (2) mentioned above) for the  $C^1$ -map  $\varphi$  of proposition 5.3 by combining the techniques of the proof of proposition 6 in [Ne] and lemma 5.5 in [Nu1]. Thus in fact, it is sufficient to assume  $F$  is  $C^2$  to guarantee the main result of the paper.

### 7.3. An *ad hoc* numerical technique

[GOY1] describes an *ad hoc* straddle technique for determining accessible periodic saddle points on the basin boundary. In [GOY1] it is assumed that there are two attractors  $A$  and  $B$ . The objective in [GOY1] is to find a saddle periodic point on the basin boundary that is accessible from  $\text{basin}(B)$ . This method worked on several test problems but had no rigorous foundation. The objective of this paper is to attack the problem raised in [GOY1] and we find a straddle method (ABST method) which has a rigorous foundation.

### 7.4. Examples

By using the SST method, in the example of the Hénon map with parameter values  $\rho = 1.812\,579\,70$ ,  $\mu = 0.022\,864\,30$  the resulting SST trajectory gives virtually the same picture as figure 4 (which was generated using the BST method). Also in this case, the ASST trajectory is similar to the ABST trajectory.

In the second Hénon example ( $\rho = 2.66$ ,  $\mu = 0.3$ ) we choose in the ABST method  $\varepsilon = 0.01$ ; the ASST method gives a similar result when  $\varepsilon = 1/30$  is chosen.

### 7.5. Smooth or fractal basin boundaries

The accessible basin boundary procedure is valid for smooth as well as fractal basin boundaries.

## References

- [AS] Alligood K T and Sauer T 1988 Rotation numbers of periodic orbits in the Hénon map *Commun. Math. Phys.* **120** 105–19
- [AY] Alligood K T and Yorke J A 1989 Accessible saddles on fractal basin boundaries *Preprint*
- [BGOYY] Battelino P M, Grebogi C, Ott E, Yorke J A and Yorke E D 1988 Multiple coexisting attractors, basin boundaries and basic sets *Physica* **32D** 296–305
- [BP] Berman A and Plemmons R J 1979 *Nonnegative Matrices in the Mathematical Sciences* (New York: Academic)

- [B] Bowen R 1975 Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms *Lecture Notes in Mathematics* **470** (Berlin: Springer)
- [B] Bowen R and Ruelle D 1975 The ergodic theory of Axiom A flows *Invent. Math.* **29** 181–202
- [GH] Guckenheimer J and Holmes P 1983 Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields *Applied Mathematical Sciences* **42** (Berlin: Springer)
- [GNOY] Grebogi C, Nusse H E, Ott E and Yorke J A 1988 Basic sets: sets determine the dimension of basin boundaries *Dynamical Systems: Proc. University of Maryland 1986–87 (Lecture Notes in Mathematics 1342)* ed J C Alexander (Berlin: Springer) pp 220–50
- [GOY1] Grebogi C, Ott E and Yorke J A 1987 Basin boundary metamorphoses: changes in accessible boundary orbits *Physica* **24D** 243–62
- [GOY2] Grebogi C, Ott E and Yorke J A 1987 Chaos, strange attractors, and fractal basin boundaries in nonlinear dynamics *Science* **238** 632–8
- [HJ] Hammel S M and Jones C K R T 1989 Jumping stable manifolds for dissipative maps of the plane *Physica* **35D** 87–106
- [KG] Kantz H and Grassberger P 1985 Repellers, semi-attractors, and long-lived chaotic transients *Physica* **17D** 75–86
- [M] de Melo W 1973 Structural stability of diffeomorphisms on two-manifolds *Invent. Math.* **21** 233–46
- [MGY] McDonald S W, Grebogi C, Ott E and Yorke J A 1985 Fractal basin boundaries *Physica* **17D** 125–53
- [NP] Newhouse S and Palis J 1973 Hyperbolic nonwandering sets on two-dimensional manifolds *Dynamical Systems* M M Peixoto (New York: Academic) pp 293–301
- [Ne] Newhouse S E 1979 The abundance of wild hyperbolic sets and non-smooth stable sets for diffeomorphisms *Publ. Math. IHES* **50** 101–51
- [Nu1] Nusse H E 1987 Asymptotically periodic behaviour in the dynamics of chaotic mappings *SIAM J. Appl. Math.* **47** 498–515
- [Nu2] Nusse H E 1988 Qualitative analysis of the dynamics and stability properties for Axiom A maps *J. Math. Anal. Appl.* **136** 74–106
- [NY1] Nusse H E and Yorke J A 1989 A procedure for finding numerical trajectories on chaotic saddles *Physica* **36D** 137–56
- [NY2] Nusse H E and Yorke J A 1991 Analysis of a procedure for finding numerical trajectories close to chaotic saddle hyperbolic sets *Ergod. Theor. Dynam. Syst.* **11** 189–208.
- [PR] Peitgen H O and Richter P H 1986 *The Beauty of Fractals* (Berlin: Springer)
- [PT] Palis J and Takens F 1987 Homoclinic bifurcations and hyperbolic dynamics, *16th Colóquio Brasileiro Matemática*, IMPA 1987
- [Y] Yorke J A 1989 *DYNAMICS. A Program for IBM PC Clones* 1989
- [YKY] You Z, Kostelich E J and Yorke J A 1991 Calculating stable and unstable manifolds *Int. J. Bifurcation and Chaos* **1** in press.

## Calculating topological entropies of chaotic dynamical systems

Qi Chen, Edward Ott<sup>1</sup> and Lyman P. Hurd

*Laboratory for Plasma Research, University of Maryland, College Park, MD 20742, USA*

Received 13 February 1991, revised manuscript received 26 March 1991, accepted for publication 3 April 1991

Communicated by A.P. Fordy

We present an efficient algorithm for calculating topological entropies of chaotic dynamical systems. The method applies to chaotic attractors as well as chaotic saddles.

The quantitative characterization of chaotic processes has proven to be an important issue in nonlinear dynamics. Calculations of Lyapunov exponents and fractal dimensions have been very useful in this regard. Another fundamental quantity is the topological entropy [1,2], which characterizes the complexity of the orbit structure of a given dynamical system. The topological entropy is invariant under topological conjugacy of the dynamical systems (i.e. it is preserved by continuous and not necessarily differentiable changes of variables).

The general definition of topological entropy is computationally unwieldy, so calculations invariably depend on theorems which give simpler or less general definitions.

For an axiom A diffeomorphism  $T$  (see ref. [3] for a definition of the conditions satisfied by an axiom A system) the topological entropy is the exponential growth rate of the number of periodic points [3]. Let  $P_n$  be the number of fixed points of the  $n$  times iterated map  $T^n$ , thus,  $P_n$  counts the number of points of period  $n$  plus the number of points whose period divides  $n$ . The topological entropy,  $h(T)$ , satisfies

$$h = \lim_{n \rightarrow \infty} \frac{\log P_n}{n}. \quad (1)$$

Thus for  $N$  sufficiently large, we have the approximation

$$h \approx \frac{\log P_N}{N}. \quad (2)$$

Chaotic systems encountered in applications are often not axiom A. Nevertheless, for non-axiom A situations, it is often assumed that eq. (2) continues to hold. Even so, chaotic systems tend to be numerically unstable, and this can make it difficult to obtain a sufficiently large number of periodic orbits to use in eq. (2). Calculations based on this method require ingenuity and have been carried out in a few cases [4,5].

Another approach due to Newhouse and Yorke relates the topological entropy to the maximum exponential growth rate of a  $k$ -dimensional volume in the phase space [1]. For two-dimensional maps, Newhouse uses these results to obtain numerical bounds on the entropy by computing the exponential growth rate of the length of a typical line segment.

Recently, a more sophisticated technique based on generating partitions of chaotic attractors has been proposed. This method seems to yield precise estimates on topological entropies. However, generating partitions are usually difficult to construct [6].

In this note, we introduce a new algorithm for calculating the topological entropy which is particularly simple and efficient, and may in some cases have some advantages over previous methods. It applies to chaotic attractors as well as chaotic saddles.

Consider an invertible map of the plane  $(x', y') = T(x, y)$ . Choose a compact volume  $V$ . Normally we choose  $V$  to contain the chaotic invariant set of

<sup>1</sup> Also at: Department of Electrical Engineering and Department of Physics.

the map. However, since the topological entropy of  $F$  is bounded below by the entropy of  $F$  restricted to any subregion, our algorithm obtains lower bounds even when this is not the case. This fact is useful if one does not know a priori bounds on the dynamics. We assume that under the action of the inverse map  $F^{-1}$ , all points in  $V$  except for a set of Lebesgue measure zero (the invariant set and its unstable manifold) eventually escape  $V$ . This is true, for example, for area contracting maps such as the dissipative Hénon map. Consider the intersection of  $V$  with its preimages,  $V_n = V \cap F^{-1}(V) \cap F^{-2}(V) \cap \dots \cap F^{-n}(V)$ . For large  $n$ ,  $V_n$  generally consists of disjoint elongated strips lying in the direction of the stable manifold of  $F$  for the invariant set contained in  $V$ . In the limit  $n \rightarrow \infty$ ,  $V_n$  is the intersection of the stable manifold with  $V$ . Let us denote the total number of disjoint components in  $V_n$  by  $N_n$  (in the case of the standard horseshoe map, this number is  $2^n$ ).

The theoretical basis for our algorithm lies in the following model situation (see ref. [7]). Let  $V$  be a rectangle whose sides are roughly parallel to the stable and unstable directions of the invariant set. If  $F(V) \cap V$  consists of  $m$  horizontal strips and  $F^{-1}(V) \cap V$  consists of  $m$  vertical strips, and  $F$  uniformly contracts horizontal strips, and  $F^{-1}$  uniformly contracts vertical strips, then  $F$  restricted to the non-empty invariant set  $\Lambda = \bigcap_{n=-\infty}^{\infty} F^n(V)$  is conjugate to the full shift on  $m$  symbols which has entropy  $\log m$  and therefore the map  $F$  has entropy at least  $\log m$  (see ref. [7] for details).

Given the region  $V$  and the map  $F$ , often the above hypotheses are not satisfied, but are satisfied by an iterate,  $F^n$  and a possibly smaller region  $V' \subseteq V$ <sup>†</sup>. Recalling that  $N_n$  is the number of disjoint strips in  $F^{-n}(V) \cap V$  the above argument implies that the entropy of  $F^n$  is at least  $\log N_n$ . Since  $h(F^n) = nh(F)$ , we define

$$s_n = \log N_{n+1} - \log N_n. \quad (3)$$

If the above hypotheses are satisfied by the region  $V$  and iterate  $n$ , the above estimate forms a rigorous lower bound. In case there is explicit checking of these hypotheses is impractical, we examine convergence

behavior of  $s_n$  for large  $n$ . Alternatively, we can plot  $\log N_n$  versus  $n$  and estimate  $h(F)$  as the slope of the fitted curve (discarding a suitable number of small  $n$  values).

To obtain an estimate of the number of disjoint strips in  $V_n$ , let  $T(x)$  denote the smallest value of  $n$  such that  $F^{-n}(x)$  is not in  $V$ . We call  $T(x)$  the inverse escape time from  $V$ . Now consider a line cutting transversely across the stable manifold. Then this line also cuts through all strips in  $V_n$  for large  $n$ , since each strip of  $V_n$  lies basically along the direction of the stable manifold. Hence,  $N_n$  is given by the number of intervals where  $T(x) \geq n$  in a typical one-dimensional line cut. In practice, we count the number of such intervals where  $T(x) \geq n$  for successively larger values of  $n$  and calculate the quantity  $s_n$  up to a certain level, or until it converges within a given tolerance. Although  $h$  obtained in this fashion only gives a lower bound for the topological entropy, for all the systems where comparisons with previous calculations are available, this algorithm appears to yield very sharp lower bounds.

We remark that in studying chaotic scattering in two-dimensional Hamiltonian flows, Kovács and Tél have obtained a similar quantity,  $K_0$ , for the Poincaré map on a surface of section. They call  $K_0$  the topological entropy of the scattering process [8]. Their method is similar to ours except that we use  $F^{-1}$  while they use  $F$  (the topological entropy of a map and its inverse are the same). Using  $F^{-1}$ , however, allows us to obtain the entropy of chaotic attractors (this is not possible using the method of ref. [8], which was designed for chaotic saddles).

We first illustrate our algorithm for the Hénon map,

$$x_{n+1} = a - x_n^2 + by_n, \quad y_{n+1} = x_n. \quad (4)$$

Set  $b=0.3$ , in the parameter range  $1.4 \leq a \leq 4.0$ , the invariant set of the Hénon map changes from a strange attractor to a strange saddle, and finally to a full 2-shift (horseshoe). For  $a$  sufficiently large, the topological entropy saturates at  $\log 2$ . It can be shown that the invariant set of the Hénon map is included in the square  $\max(|x|, |y|) \leq R$ , where [9]

$$R = \frac{1}{2} \{1 + |b| + [(1 + |b|)^2 + 4a]^{1/2}\}.$$

This is the region  $V$  which we use for calculating the inverse escape time function. For simplicity, we take

<sup>†</sup> Recall that the topological entropy of  $F$  restricted to  $V'$  gives a lower bound for the topological entropy of  $F$  restricted to  $V$ .

a vertical one-dimensional line through the origin  $x=0, y=0$  and calculate  $T(x)$  at regularly spaced intervals. This is shown in fig. 1 for  $a=3.0$ , where the invariant set is topologically a full 2-shift (horseshoe). There is a natural Cantor set level structure in the inverse escape time function. At level 1, there are two intervals from which it requires at least two backward iterations to escape the square  $V$ ; at level 2, there are four intervals from which it requires at least three backward iterations to escape  $V$ , etc. The intersection of these intervals is the intersection of the stable manifold of the invariant set with the vertical axis.

Using a double-precision algorithm, we are able to calculate the inverse escape time function up to level 20. The algorithm is implemented as follows. Starting from the initial interval  $I_0$  given by the intersection of the vertical axis with  $V$ , we interpolate  $I_0$  with a uniform grid of  $N=50$  points and calculate the inverse escape time for each point with cutoff time  $n=2$ . We find all the intervals  $I_1$ 's in the grid where the inverse escape time function is greater than 1. We then interpolate again each interval  $I_1$  with 50 points, calculate the inverse escape time for each point with cutoff time  $n=3$ , and find all the subintervals  $I_2$ 's where the inverse escape time function is greater than 2, etc. Assuming each iteration of the Hénon map costs about 10 machine instructions and the topological entropy to be calculated is  $\log 2$ , the whole calculation up to level 15 then costs approx-

imately 32 million machine instructions. On a 10 MIPS workstation, the whole computation takes approximately 3 s. We can achieve better precision by going to higher levels or interpolating more points in the grid. The calculation time typically increases with the level at an exponential rate given by the topological entropy. Usually, level 10 calculation (1 million machine instructions, or 0.1 s on a 10 MIPS workstation) yields good estimates on the entropy for chaotic systems. (For instance, for the Hénon attractor at  $a=1.4, b=0.3$ , level 10 calculation gives  $s=0.660$ , while level 15 gives  $s=0.670$ , a relative error of less than 2%. Note this value is consistent with the one obtained in ref. [6].) In all our numerical examples, the logarithms are taken to be base 2.

Fig. 2 shows the topological entropy for the Hénon map at  $b=0.3$  in the parameter range  $1.4 \leq a \leq 3.0$ . It is calculated with 100 interpolation points at level 15. This figure seems to be identical (with better precision) with the one obtained by Biham and Wenzel [4]. Note that there are plateau regions where the entropy is constant. This is because for any parameter value where the invariant set is hyperbolic, the topological entropy must be locally constant due to the structural stability of hyperbolic sets. The whole calculation with 260 parameter values takes about 50 min on a 10 MIPS workstation.

We also apply our method to open Hamiltonian systems. Generically, the phase space of Hamilto-

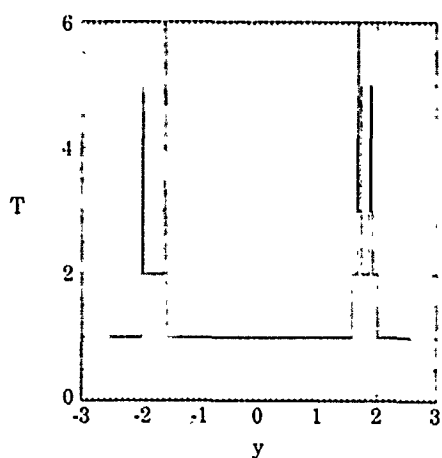


Fig. 1. The inverse escape time function for the Hénon map at  $a=3.0, b=0.3$  for a vertical cut through the origin  $x=0, y=0$ .

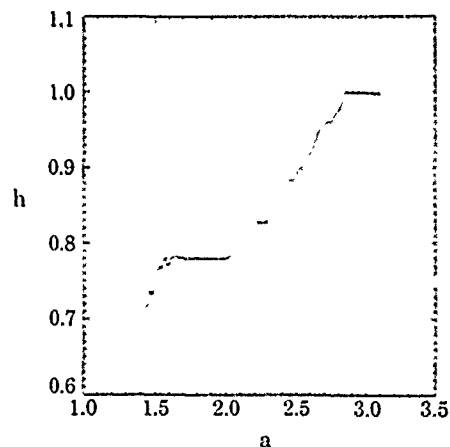


Fig. 2. The topological entropy for the Hénon map as a function of  $a$  at  $b=0.3$ . This graph is obtained using the method described in the text at level 15 with 100 interpolation points.

nian systems has mixed components [10]: regular rotational motions of KAM type, and irregular motions with positive Lyapunov exponents. If the irregular component is noncompact, its only bounded invariant subsets are strange saddles. The topological entropy is related to the escape dynamics from the saddle [5]. We wish to calculate the topological entropies of such systems. One example is given by the area-preserving sawtooth map on the plane [5],

$$y_{n+1} = y_n + Kf(x_n), \quad x_{n+1} = x_n + y_{n+1}. \quad (5)$$

Here  $f(x)$  is a sawtooth function,

$$f(x) = x - 0.5 - [x], \quad (6)$$

where  $[x]$  denotes the greatest integer in  $x$ . Note that  $f(x)$  is discontinuous on the line  $x=0$ , therefore the sawtooth map is piecewise linear with constant Jacobian matrix except on the line  $x=0$ . The nonlinearity of the map comes from this line of discontinuity. For  $K>0$ , the map is uniformly hyperbolic except on the discontinuity line, hence, there are no KAM curves in the phase space. The Lyapunov number  $\Lambda$  is related to the parameter  $K$  by

$$\Lambda = 1 + 0.5[K + (K^2 + 4K)^{1/2}].$$

Under the action of the sawtooth map, almost all initial conditions inside the fundamental region  $V = \{x: |x| \leq 0.5\}$  escape to infinity. It can be shown that there is an unstable invariant set  $\Gamma$  in  $V$  [5]. Fig. 3 shows this invariant set at  $\Lambda = 2.4$ . We will calculate the topological entropy for this invariant set as a

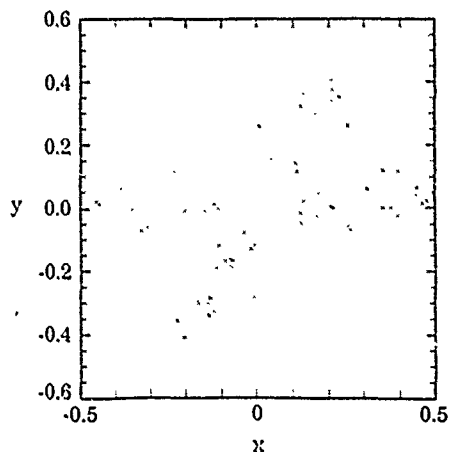


Fig. 3. The unstable invariant set for the sawtooth map at  $\Lambda = 2.4$ .

function of  $\Lambda$ . (We note that the topological entropy of the sawtooth map defined on the plane is different from the topological entropy of the same map defined on the torus. In the latter case, the space is compact, the chaotic invariant set is the whole torus, which contains  $\Gamma$  as a subset. The topological entropy of the latter is given by the Lyapunov exponent  $\log \Lambda$ , the uniform expansion rate of a line segment.) When  $\Lambda > 3$ , we can show that the invariant set is a full 2-shift [5], therefore, the topological entropy saturates at  $\log 2$  when  $\Lambda > 3$ .

For convenience, we choose the cut at  $x = -0.5$ . The topological entropy is shown in fig. 4 for  $2 \leq \Lambda \leq 3$ . The solid curve is the entropy obtained by counting the number of periodic points of the map by using the coding scheme of ref. [5], the dots are entropies calculated with our algorithm at level 18. The agreement is excellent. When  $\Lambda < 2$ , the convergence for both methods becomes slow, and we find it prohibitive to obtain the entropy value without going to a higher precision algorithm. We note that there is no apparent plateau structure in fig. 4. This is because the invariant set is not everywhere hyperbolic in this parameter range.

We have also calculated the topological entropy for the corresponding invariant set of the standard map on the plane. The standard map is given by replacing the impulse function in (5) with a sinusoidal function [10].

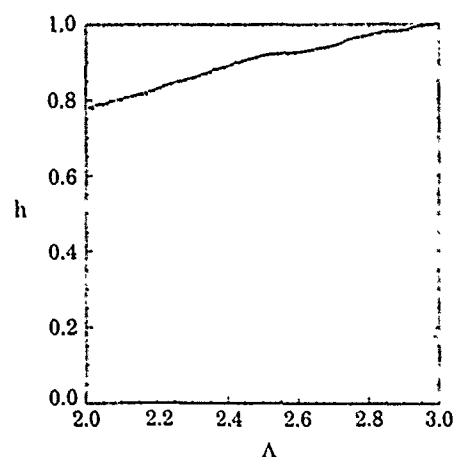


Fig. 4. The topological entropy as a function of  $\Lambda$  for the sawtooth map. The solid curve is the entropy obtained by counting the number of  $n$ -cycle fixed points, the dots are the level 18 calculations with 100 interpolation points.

$$f(x) \equiv -\sin(2\pi x)/2\pi. \quad (7)$$

For moderately large values of  $K$  (of order 1), the motion in the phase space has both regular and irregular components. However, when the parameter  $K$  is large, the map is almost hyperbolic [10]. Therefore, the invariant set contained in the fundamental region  $V = \{x: |x| \leq 0.5\}$  is a strange saddle for large  $K$ . In fig. 5, we show the topological entropy in the parameter range  $5.0 \leq K \leq 9.0$  calculated using our algorithm at level 10 (again logarithms are calculated in base 2). We see that at  $K \approx 8.4$ , the topological entropy saturates at  $\log 3$ , indicating the invariant set is topologically a 3-shift. Indeed, this is the typical dynamics of the standard map for large parameter  $K$ . We again note that in the entropy function there are obvious plateau regions where the topological entropy remains constant, similar to the case of the Hénon map.

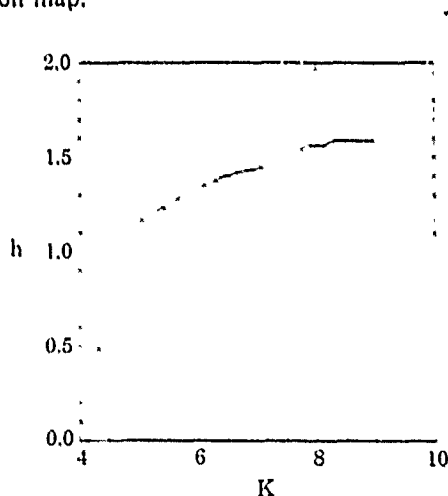


Fig. 5. The topological entropy for the standard map as a function of  $K$ . 100 interpolation points at level 10.

In conclusion, we have presented an efficient algorithm for calculating the topological entropy of chaotic dynamical systems.

QC wants to thank the Aspen Center for Physics for its hospitality and Rex Skodje for discussions. This work was supported by the Office of Naval Research (Physics), by the Department of Energy (Scientific Computing Staff Office of Energy Research) and by the Defense Advanced Research Projects Agency.

## References

- [1] S. Newhouse, Entropy and volume as measures of orbit complexity, in: *Lecture notes in physics*, Vol. 278, The physics of phase space (Springer, Berlin, 1986) p. 2.
- [2] P. Walters, *An introduction to ergodic theory* (Springer Berlin, 1982);  
N.F.G. Martin and J.W. England, *Mathematical theory of entropy* (Addison-Wesley, Reading, 1981).
- [3] R. Bowen, *Am. Math. Soc.* 154 (1971) 377;  
A.B. Katok, *Publ. Math. IHES* 51 (1980) 137.
- [4] O. Biham and W. Wenzel, *Phys. Rev. Lett.* 63 (1989) 819;  
P. Grassberger, H. Kantz and U. Moenig, *J. Phys. A* 22 (1989) 5217.
- [5] Q. Chen, I. Dana, J.D. Meiss, N. Murray and I.C. Percival, *Physica D* 67 (1990) 217;  
N. Bird and F. Vivaldi, *Physica D* 30 (1988) 164;  
I.C. Percival and F. Vivaldi, *Physica D* 25 (1987) 105.
- [6] G. D'Alessandro, P. Grassberger, S. Isola and A. Politi, *J. Phys. A* 23 (1990) 5285.
- [7] J. Guckenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems and bifurcations of vector fields* (Springer, Berlin, 1983).
- [8] Z. Kovács and T. Tél, *Phys. Rev. Lett.* 64 (1990) 1617.
- [9] R. Devaney and Z. Nitecki, *Commun. Math. Phys.* 67 (1979) 137.
- [10] B.V. Chirikov, *Phys. Rep.* 52 (1979) 262.

# On the Tendency Toward Ergodicity with Increasing Number of Degrees of Freedom in Hamiltonian Systems

Lyman Hurd

Iterated Systems, Corp.

5550A Peach Tree Parkway, Suite 545

Norcross, GA 30092

and

Celso Grebogi<sup>a</sup> and Edward Ott<sup>b</sup>

Laboratory for Plasma Research

University of Maryland

College Park, Maryland 20742-3511

## ABSTRACT

Numerical experiments on a symplectic coupled map system are performed to investigate the tendency for global ergodic behavior of typical Hamiltonian systems as the number of degrees of freedom  $N$  is increased. As  $N$  increases, we find that the fraction of phase space volume occupied by invariant tori decreases strongly. Nevertheless, due to observed very long time correlated behavior, a conclusion of effective gross ergodicity cannot be confirmed, even though extremely long numerical runs were employed.

a. and Department of Mathematics, and Institute for Physical Science and Technology.

b. and Department of Physics and Astronomy, and Department of Electrical Engineering.

The basic assumption in statistical mechanics is that of ergodicity over the phase space hypersurface determined by the global constants of the motion (e.g., total energy, total angular momentum, etc.). On the other hand, studies of Hamiltonian systems with few degrees of freedom (e.g., two) typically reveal the presence of invariant KAM tori in addition to chaotic orbits; and the existence of KAM tori yields motion that is grossly different from that assumed in statistical mechanics. A natural supposition reconciling the above contradictory views might be that, as the number of degrees of freedom is increased, the tendency for global ergodicity increases. By "tendency for global ergodicity" we mean that, for systems with many degrees of freedom (the situation of interest in statistical mechanics), the overwhelming majority of initial conditions would be ergodic over effectively all of the area of the phase space hypersurface determined by the global constants of the motion.

The purpose of this paper is to present numerical experiments which attempt to test this supposition in a specific case. In particular, we study a symplectic map system (the symplectic condition insures that the dynamics is Hamiltonian). A closely related work is that of Falcioni et al.<sup>1</sup> For other previous relevant works on Hamiltonian dynamics in higher number of degree of freedom systems see Kaneko and Bagley,<sup>2</sup> Gyorgyi et al.,<sup>3</sup> and the discussion and references in the book by Lichtenberg and Lieberman.<sup>4</sup> The main result of the present paper is that, for the system we study, the fraction of orbits on tori decreases very strongly as the number of degrees of freedom is increased, but there is still no conclusive evidence for effectively complete global ergodicity even over the very long times investigated in our numerical experiments. The latter is due to the extremely long time-scales, insensitive to machine precision, observed in the numerical experiments.

The system we studied derives from the standard map,

$$\begin{aligned}x' &= x + y, \\y' &= y + k \sin x'.\end{aligned}\tag{1}$$

In these coordinates the map can be considered as a map of the two-torus  $T^2$ ,  $0 \leq x < 2\pi$  and  $0 \leq y < 2\pi$ .

Given a positive integer  $N$ , consider the space  $(T^2)^N$  thought of as  $2n$ -tuples  $(x_0, y_0, x_1, y_1, \dots, x_{N-1}, y_{N-1})$ . We define a coupled standard map allowing symmetric

bidirectional nearest neighbor interactions,

$$\begin{aligned} x'_i &= x_i + y_i, \\ y'_i &= y_i + K \sin x'_i + CK \sin(x'_i - x'_{i-1}) + CK \sin(x'_i - x'_{i+1}), \end{aligned} \quad (2)$$

where the indices are taken modulo  $N$  and  $x_i, y_i$  are taken modulo  $2\pi$ . Here  $C$  is the coupling parameter to nearest neighbors. Letting  $K = k/(2C + 1)$ , Eqs. (2) reduce to Eqs. (1) for  $N = 1$ . We call  $k$  the nonlinearity parameter. This map is symplectic since it can be obtained from the generating function,

$$F(x, x') = \frac{1}{2} \sum_{i=1}^N (x'_i - x_i)^2 + K \cos x'_i + CK \cos(x'_i - x'_{i+1}). \quad (3)$$

One checks readily that  $y_i = \partial F / \partial x_i$ ,  $y'_i = -\partial F / \partial x'_i$ .

The original aim of our numerical experiments was the exploration of the relative measure of KAM tori as a function of the number of coupled maps. To this end, we first note that motion on KAM surfaces is quasiperiodic with all Lyapunov exponents zero, while motion not on KAM surfaces typically is chaotic and has at least one positive Lyapunov exponent. Thus we proceed as follows (see also Ref. 1). A cutoff value  $\epsilon$  for an orbit to be considered quasiperiodic was set and the number of initial conditions with largest Lyapunov exponent (LE) less than  $\epsilon$  counted. The run consisted of taking  $m$  initial conditions uniformly distributed in the  $2N$ -torus and iterating them approximately  $10^6$  times along with their tangent vectors to compute their LE's. A cutoff value  $\epsilon = 0.005$  for the largest LE was set below which an orbit was considered quasiperiodic, and the ratio of the number of quasiperiodic initial conditions to the total number of initial conditions was returned.

When the coupling coefficient is zero, the volume of the KAM tori decays exponentially with  $N$ . In particular, if  $f$  denotes the fraction of phase space occupied by KAM tori for a single standard map, Eq. (1), then the fraction of the phase space  $(T^2)^N$  for  $N$  uncoupled maps for which motion in the  $2N$  variables  $(x_0, y_0, \dots, x_{N-1}, y_{N-1})$  is quasiperiodic is  $f^N$ . When  $C > 0$ , the rate of decay was observed to increase dramatically. Results for the parameter values  $C = 0.5$ ,  $k = 0.3$  are displayed in Table 1. In this table the estimated measure of quasiperiodic (QP) initial conditions (second column) is the fraction of 8192

randomly chosen initial conditions yielding LE's less than  $\epsilon$ .

Maps	Estimated Measure of QP Initial Conditions	Iterations
1	1.000	$10^6$
2	0.403	$3.25 \times 10^6$
3	0.048	$3.25 \times 10^6$
4	0.002	$3.25 \times 10^6$
5	0.000	$1.25 \times 10^6$
6	0.000	$10^6$
7	0.000	$10^6$

Table 1: Fraction of Initial Conditions Yielding Quasiperiodic (QP) Orbits

Figures 1 show histograms of the observed distribution of maximum LE's for the 8192 randomly chosen initial conditions for  $N = 1, 2, 3, \dots, 7$  coupled maps.

The case of three maps is presented twice with different numbers of iterations for the same set of data. The observed peaks get sharper but the effect is very slow.

In most cases the following phenomena were noted:

1. The number of initial conditions following within the  $\epsilon$  bound for quasiperiodicity decreases rapidly as the number of maps increases.
2. The observed peaks grew sharper with repeated iteration—but very slowly.
3. The histograms with more than one peak preserved those peaks and they individually got sharper.

These observations might lead one to conjecture that each peak represents a distinct ergodic component with its own maximum LE.

We now discuss the behavior of six individual orbits for the  $N = 3$  case, where the orbits are chosen so that their maximum LE's calculated after  $3.25 \times 10^6$  lay in distinct regions of interest of the histogram in Fig. 1(c). The calculated LE values for these six orbits are indicated by the arrows labeled with the letters (a)–(f) along the axis of Fig.

1(c). The projection of these orbits onto the first two components ( $x_0, y_0$ ) are plotted for  $10^4$  iterations in Figs. 2(a)-(f).

Distinct orbits appeared to stay constrained in a fixed region of phase space, and this was also true when time series of  $10^5$  iterations were plotted.

Lyapunov exponents were then computed for each of these orbits for a much greater period of time ( $3 \times 10^8$  iterations). The results are shown in Fig. 3 where the letters (b)-(f) labeling the curves correspond to the orbits shown in Figs. 2(b)-(f) and the arrows shown along the horizontal axis of Fig. 1(c). The first initial condition, which was presumed quasiperiodic, remained stable during the whole process, and in fact its computed LE reached zero to machine precision. Initial condition (f) also remained at a highly stable value. The remaining four, however, appear to have started to converge slowly to a new common value.

Figures 4 break down the curves in Fig. 3 [plus orbit (a)] giving the cumulative LE and a "local" LE which is calculated in 500,000 iterate bursts. Observe the great stability of initial conditions (a) and (f).

Further studies were conducted for a variety of initial conditions and various behaviors were observed.

1. Some initial conditions appeared to lead to orbits whose LE's showed a great deal of stability (they remained essentially unchanged over the observed time scale).
2. Some initial conditions showed a high degree of stability at one value of the maximum LE but then "leaked" into a regime with a different LE.
3. Some initial conditions alternated between chaotic behavior and behavior very close to quasiperiodic.

One effect of these observations was to call into question the reliability of the LE calculations in general. Many of these calculations seemed to be stable for greater than  $10^6$  iterates before changing value. Given the relative rarity of these "leaks," it was impractical to assign any numerical value to this diffusion.

The histogram calculations were performed on a Connection Machine using (of necessity) single-precision arithmetic. The orbit calculations were performed on a DecStation 3100 using double precision. To examine the effect of machine precision several of the long-term LE calculations were done at both single and double precision. The observed behavior was qualitatively the same; the observed leakage between regions of different LE occurred in each case (at slightly different iterates).

This work was supported by the Office of Naval Research (Physics Branch), by the Department of Energy (Scientific Computing Staff, Office of Energy Research), and by the Defense Advanced Research Projects Agency.

## REFERENCES

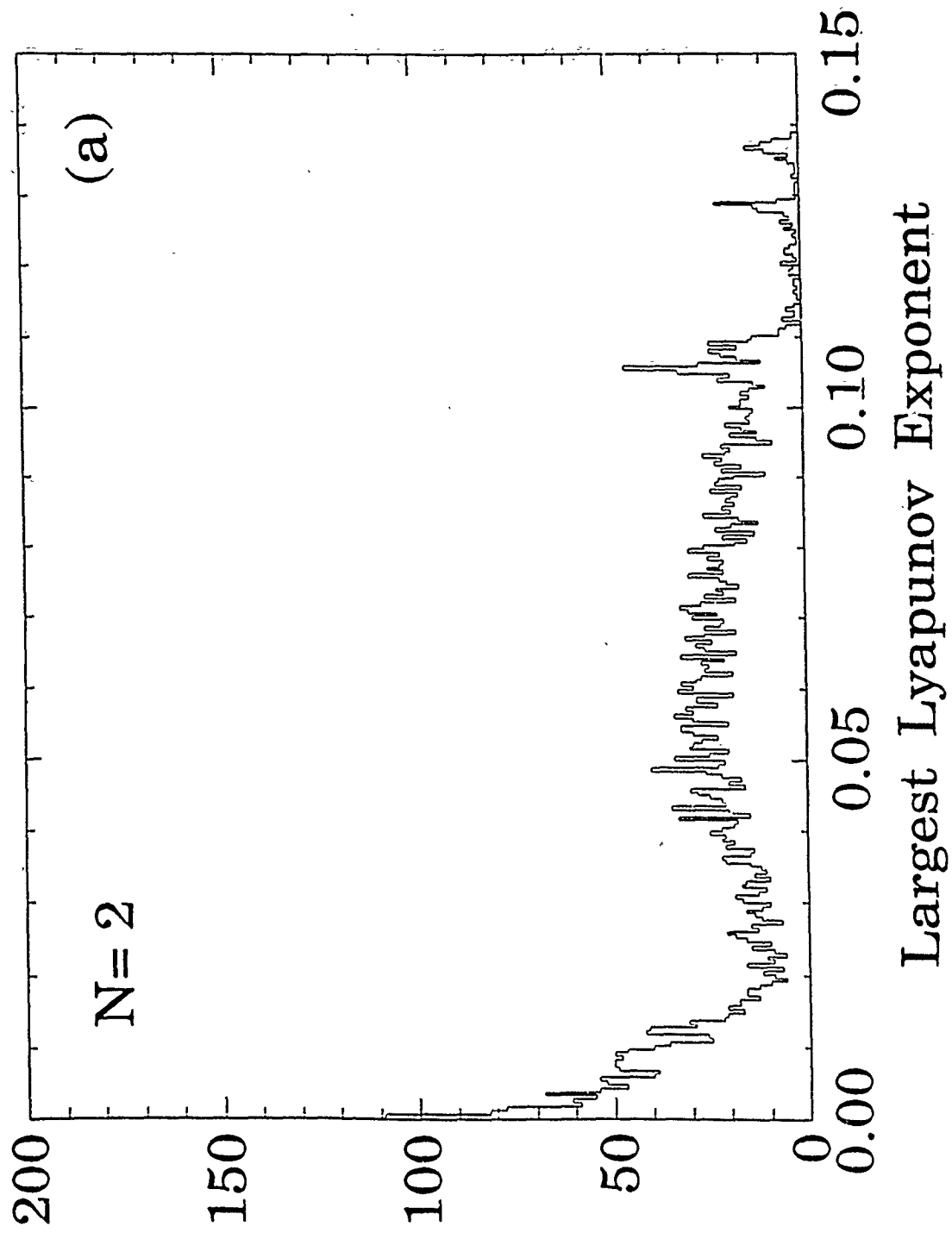
- [1. ] M. Falcioni, U. Marini Bettolo Marconi, A. Vulpiani, Phys. Rev. A **44**, 2263 (1991).

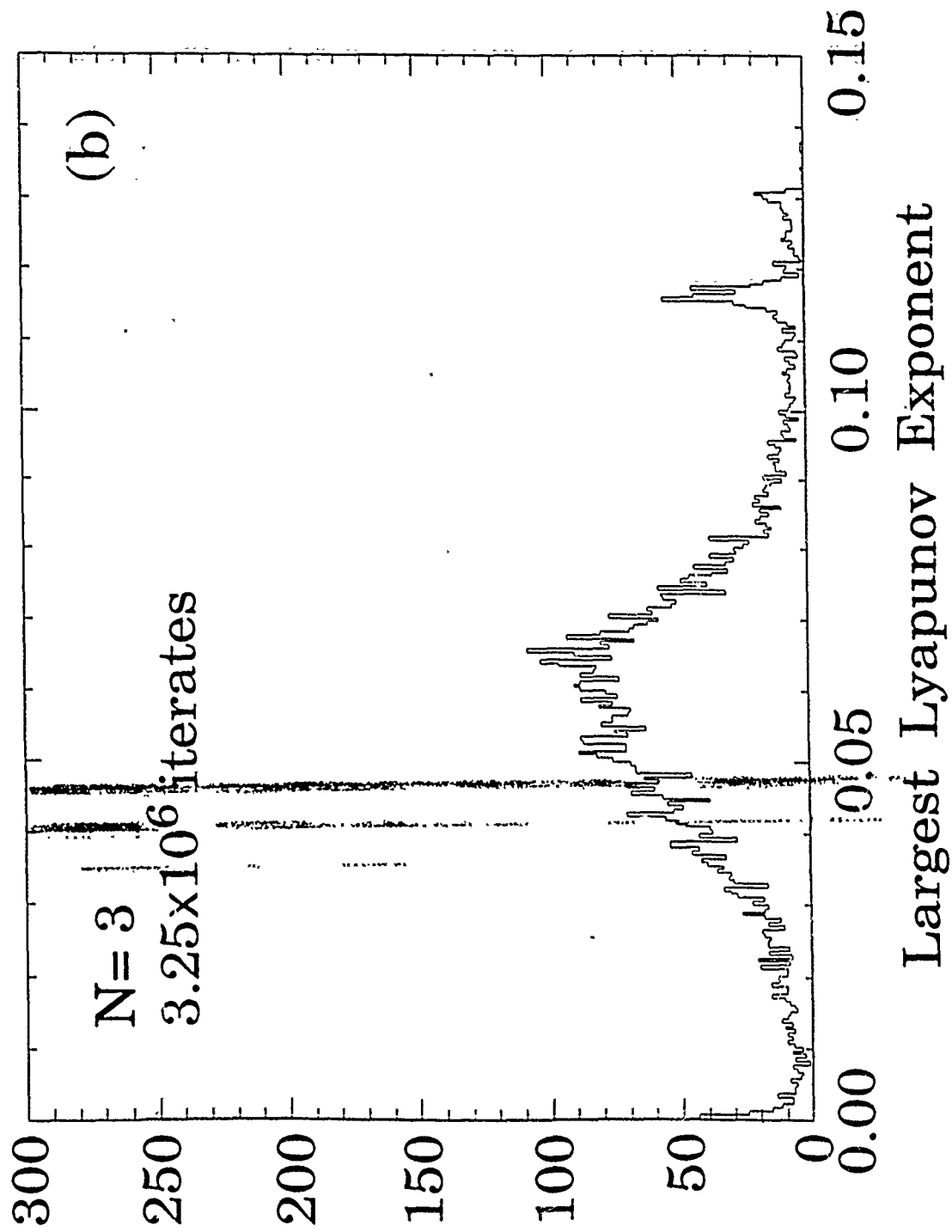
We had completed our research at the time of the publication of the paper of Falcioni et al. Because of the similarity of that work and ours, in this Letter we shall be somewhat briefer than we otherwise might have been, and will also emphasize that part of our work which is different from that of Falcioni et al.

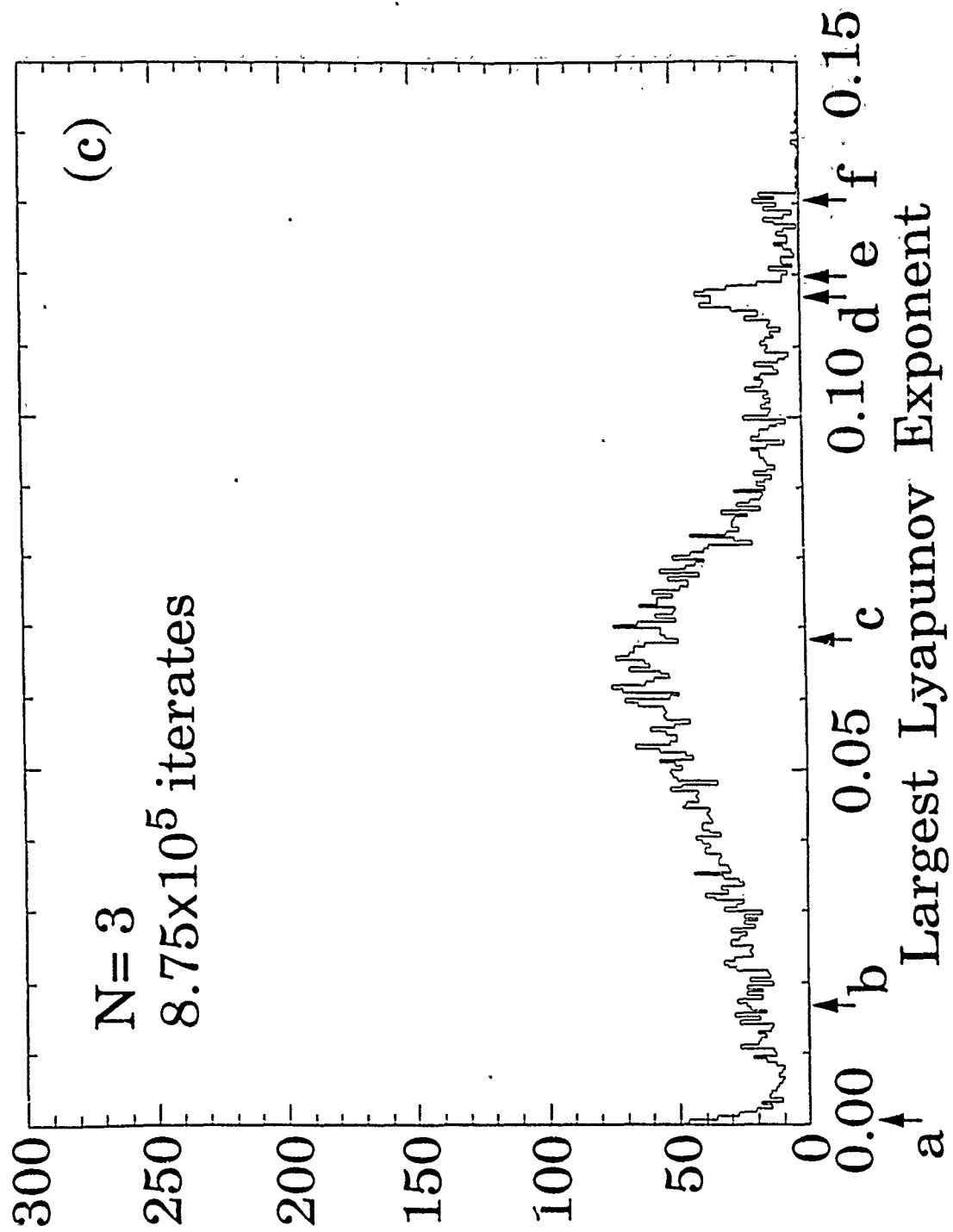
- [2. ] K. Kaneko and R. J. Bagley, Phys. Lett. A **110**, 435 (1985).
- [3. ] G. Gyorgyi, F. A. Ling and G. Schmidt, Phys. Rev. A **40**, 5311 (1989).
- [4. ] A. J. Lichtenberg and M. A. Lieberman, *Regular and Stochastic Motion* (Springer-Verlag, Berlin, 1983).

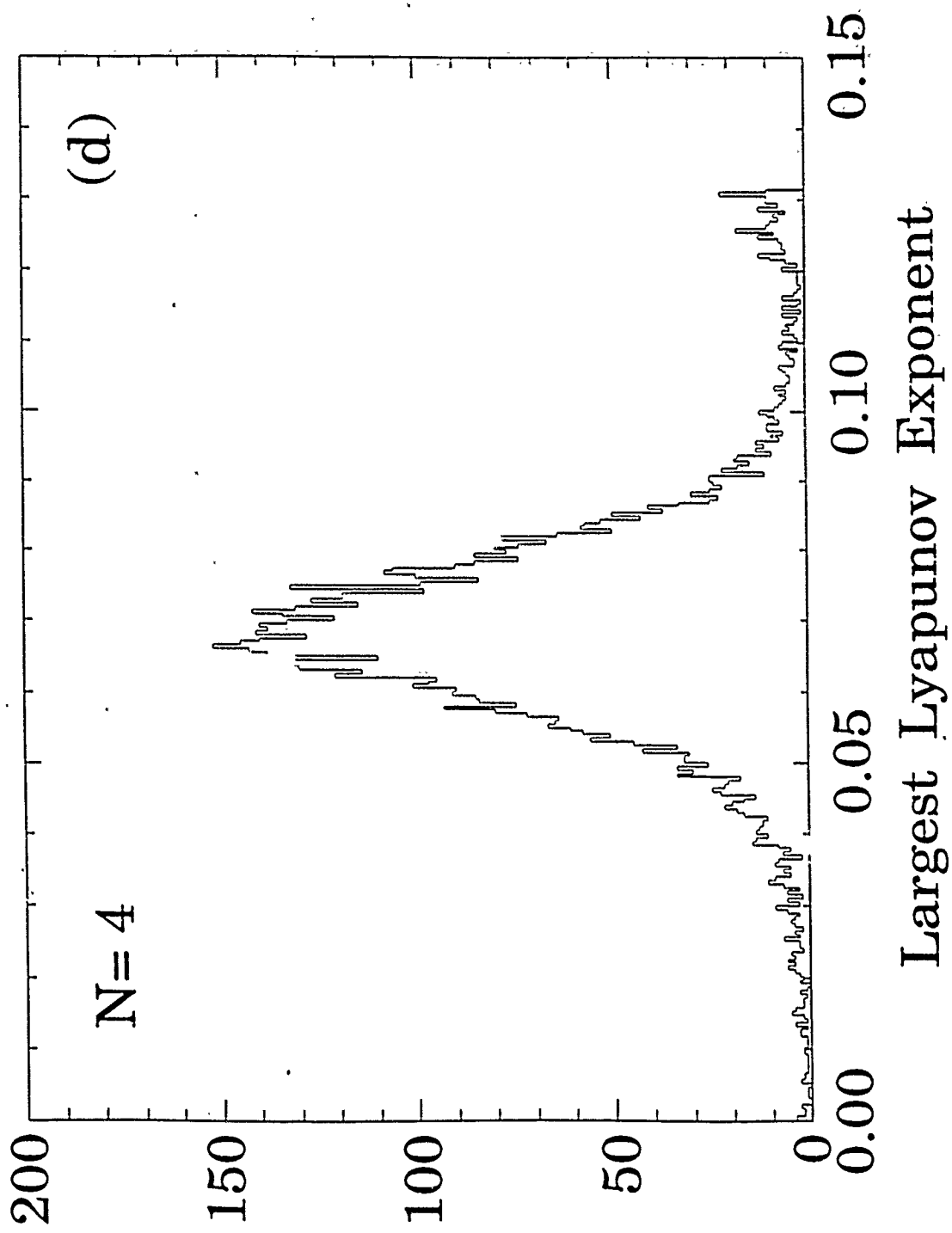
## FIGURE CAPTIONS

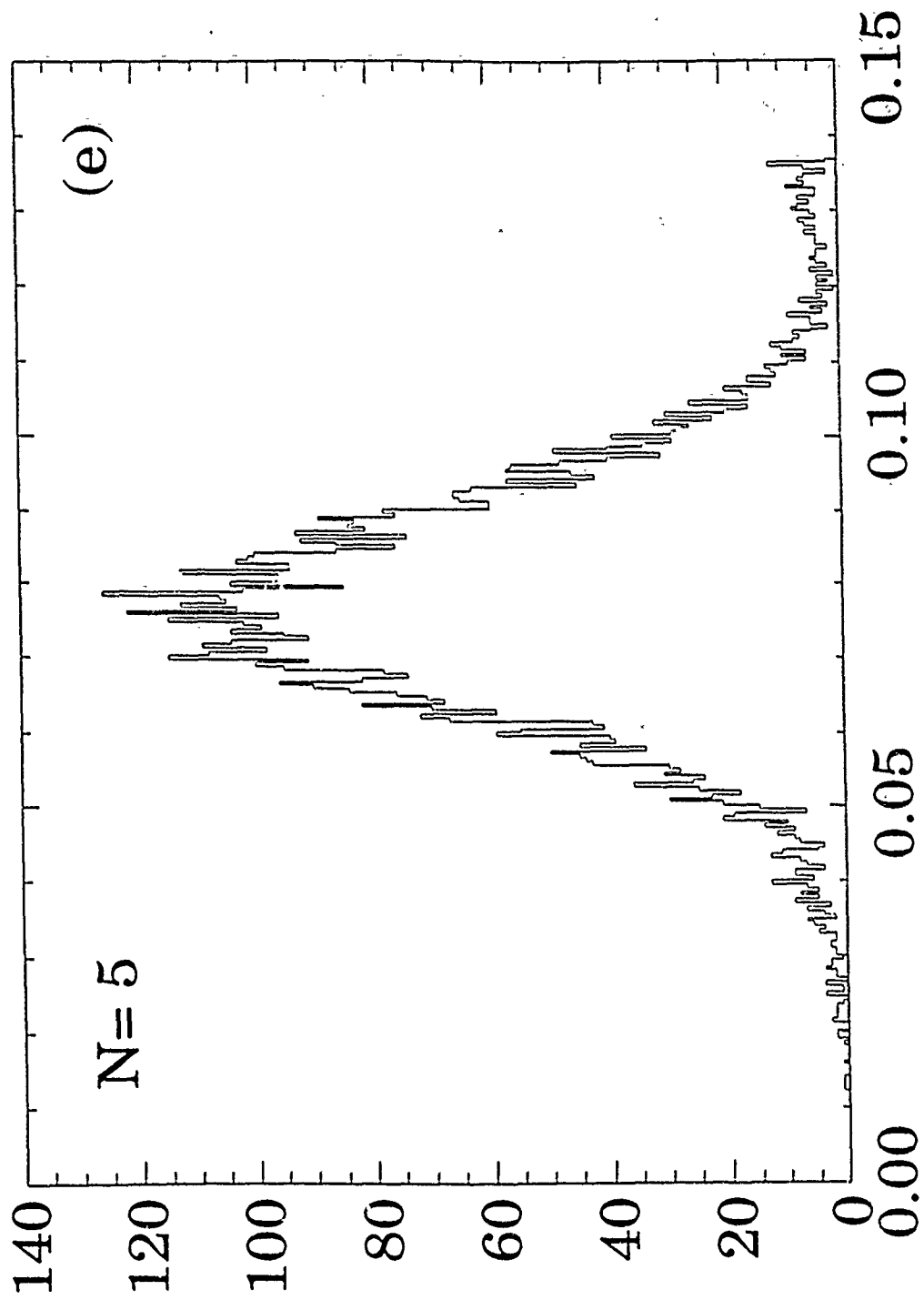
1. Histograms of maximum Lyapunov exponents for 8192 initial conditions and  $N = 2, 3, \dots, 7$ . For (a) the value in the first histogram bin is about 2700, far off the scale shown. In general, the value in the first bin is an estimate of the number of quasiperiodic orbits.
2. Projection of six individual orbits for  $N = 3$  onto the plane corresponding to the first two components.  $10^4$  iterations are plotted. The calculated maximum LE's for these orbits are (a)  $8 \times 10^{-5}$  (quasiperiodic), (b) 0.0166, (c) 0.0676 [corresponding to the lower LE peak in Fig. 1(c)], (d) 0.1170 [corresponding to the higher LE peak in Fig. 1(c)], (e) 0.1191, and (f) 0.1300. These LE values are indicated along the horizontal axis of Fig. 1(c).
3. Maximum calculated LE as a function of the number of map iterations for the five orbits corresponding to Figs. 2(b)–2(f).
4. Cumulative and “local” maximum LE for the orbits corresponding to Fig. 3.

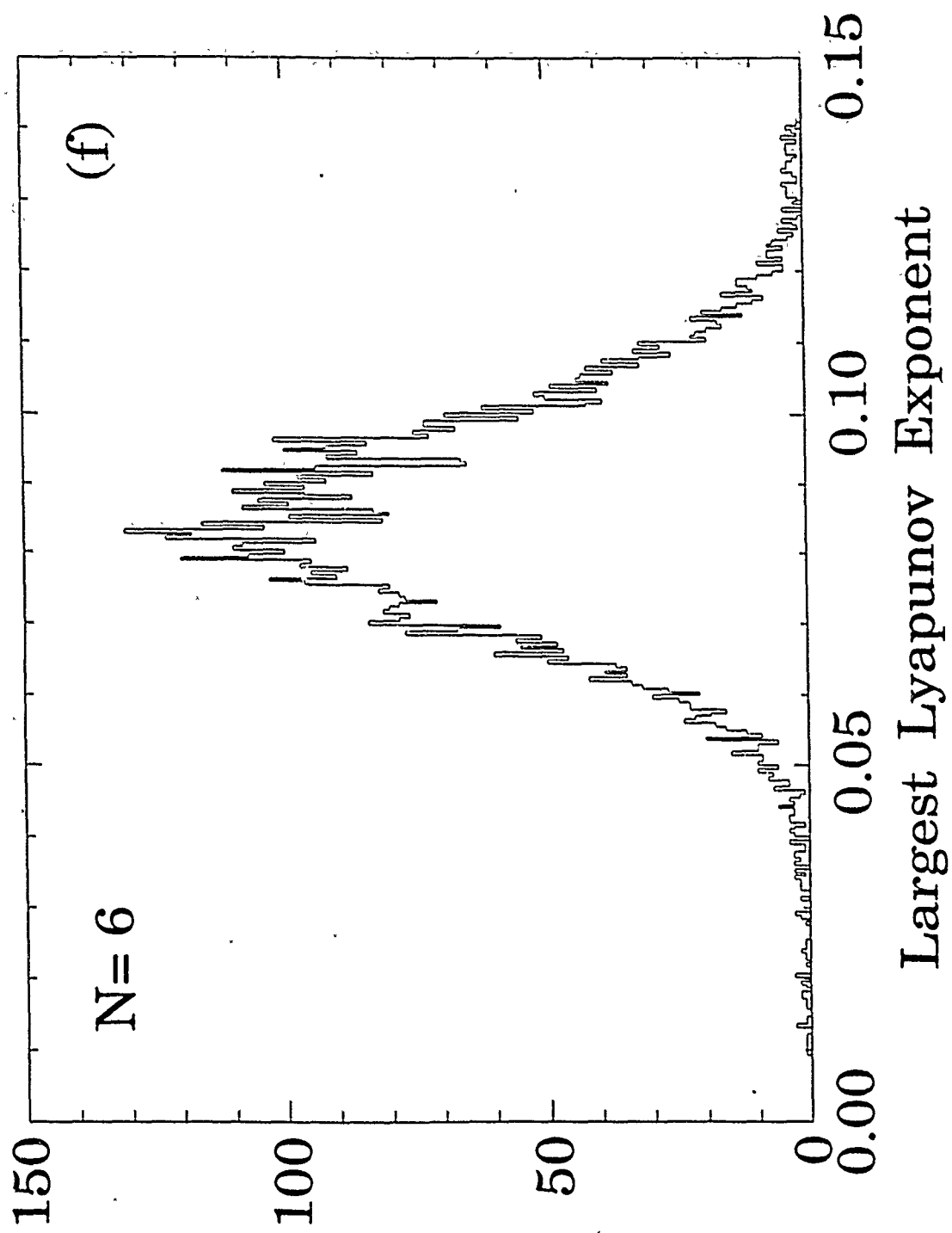


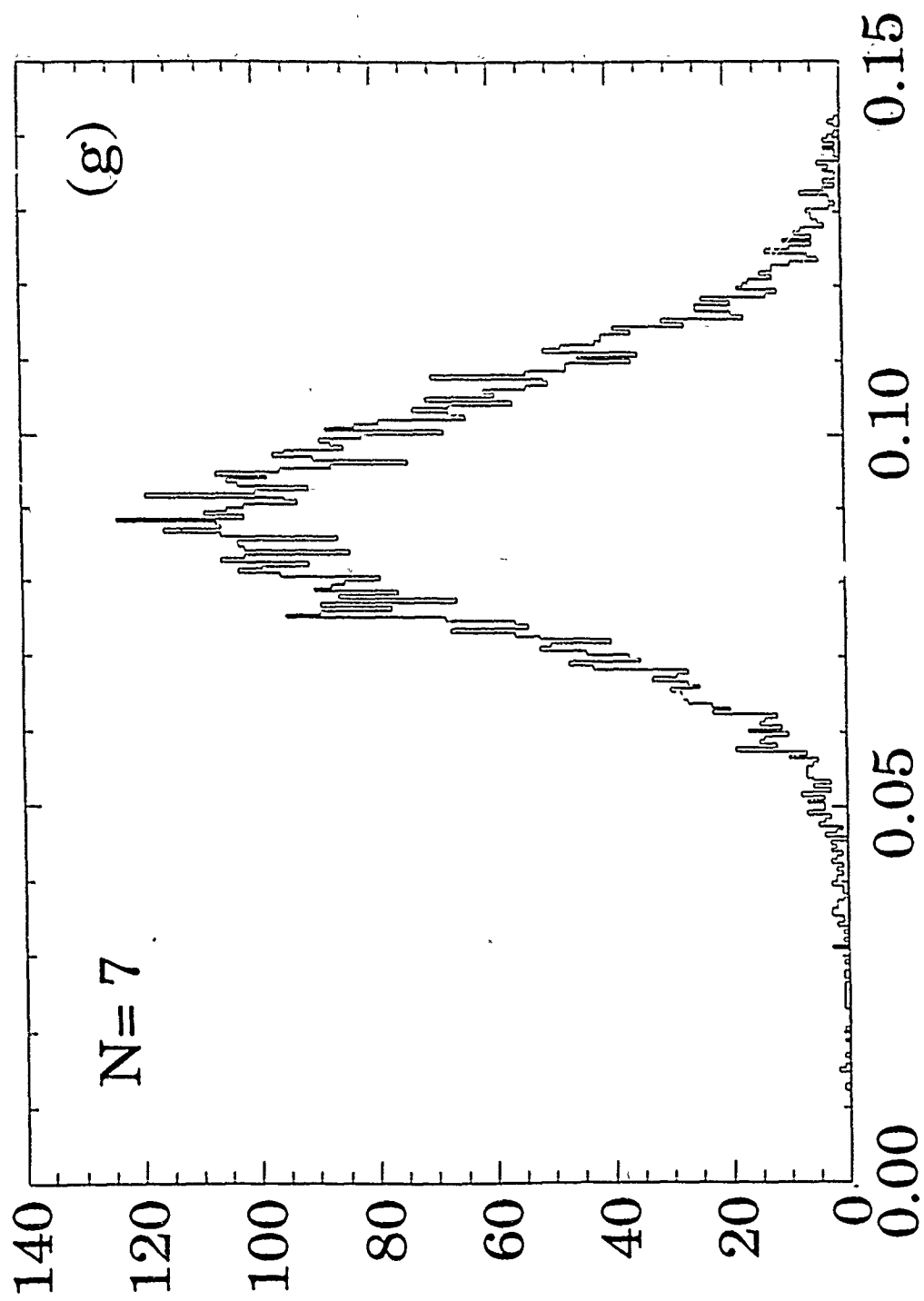


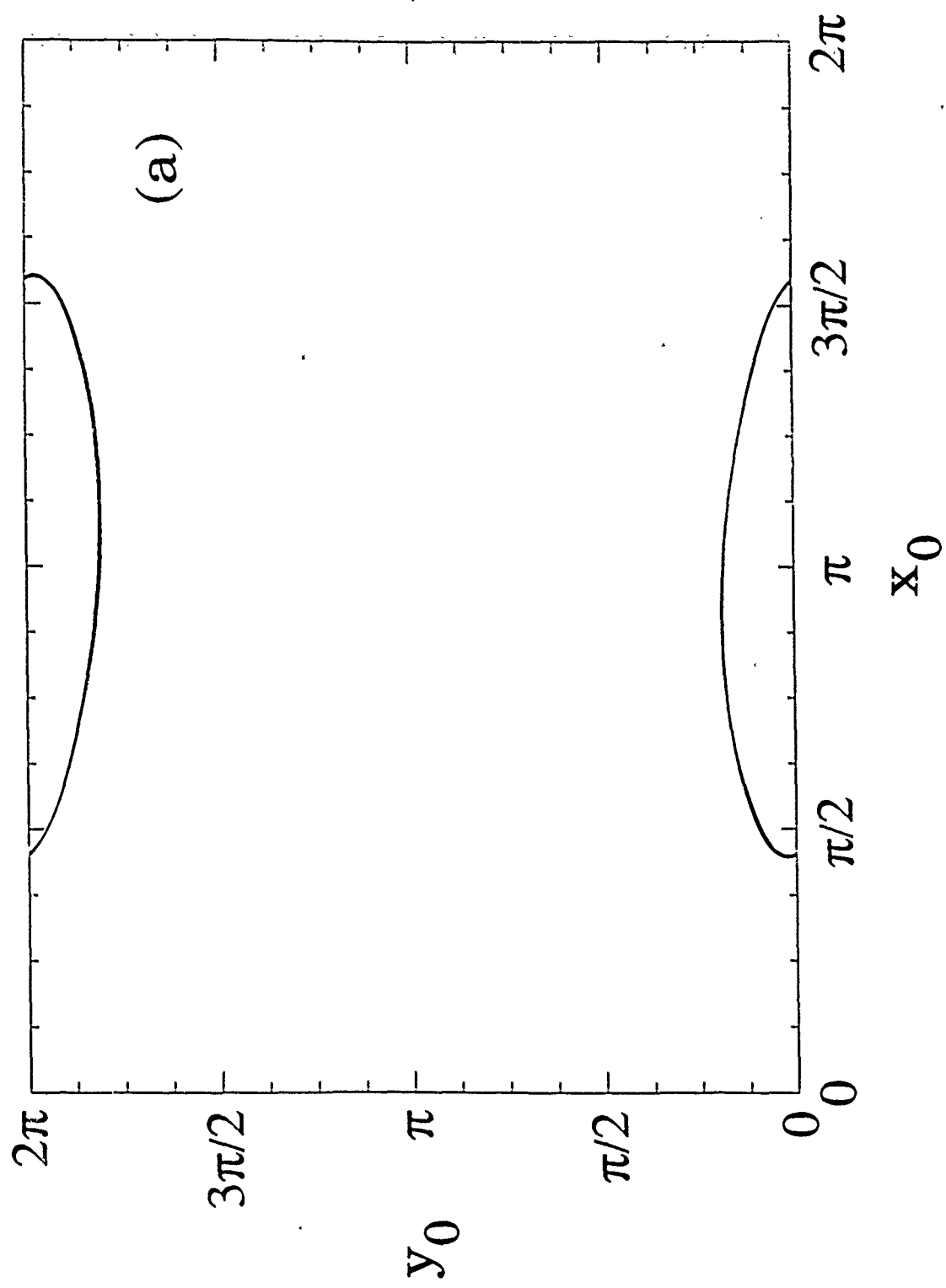


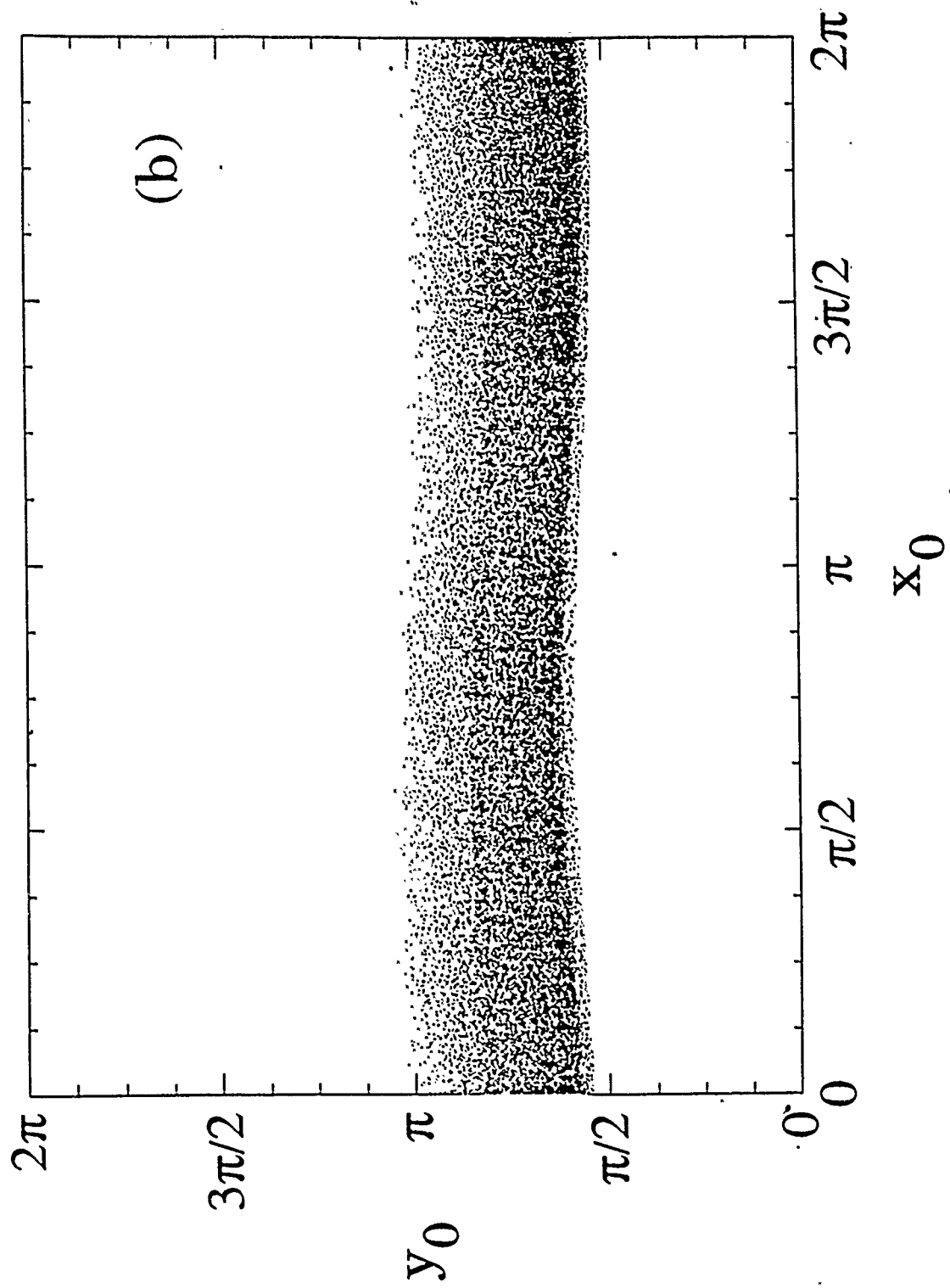


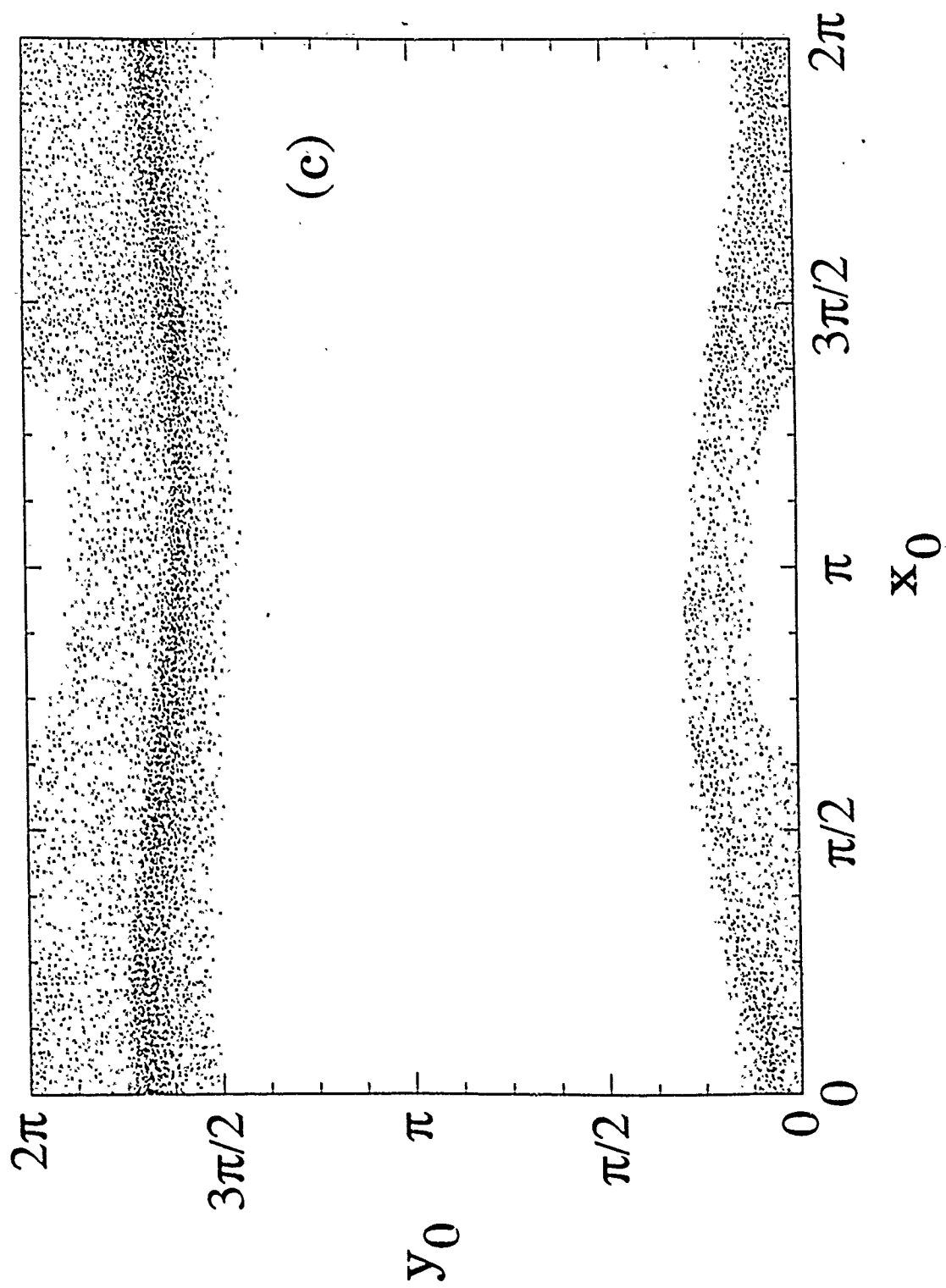


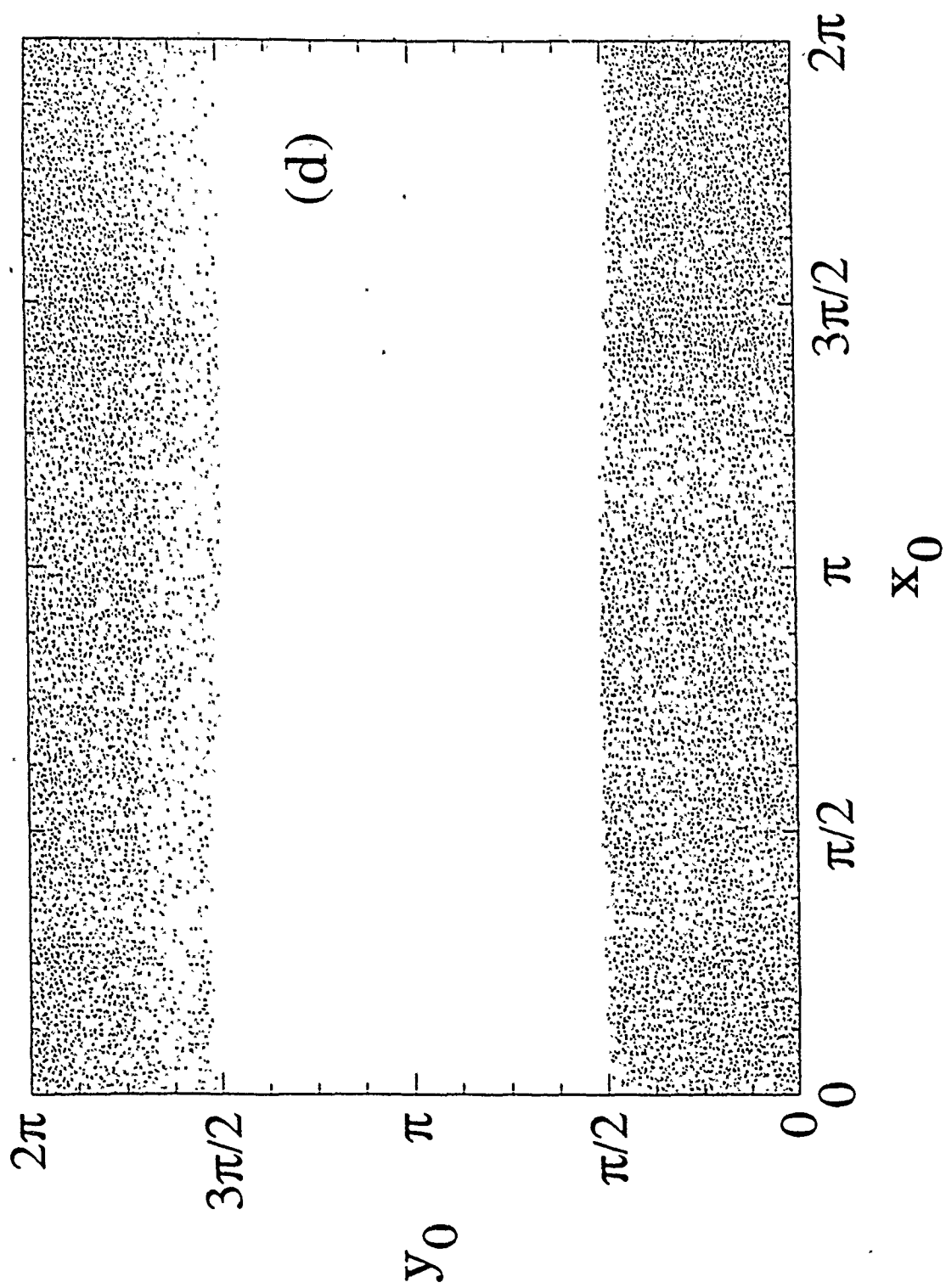


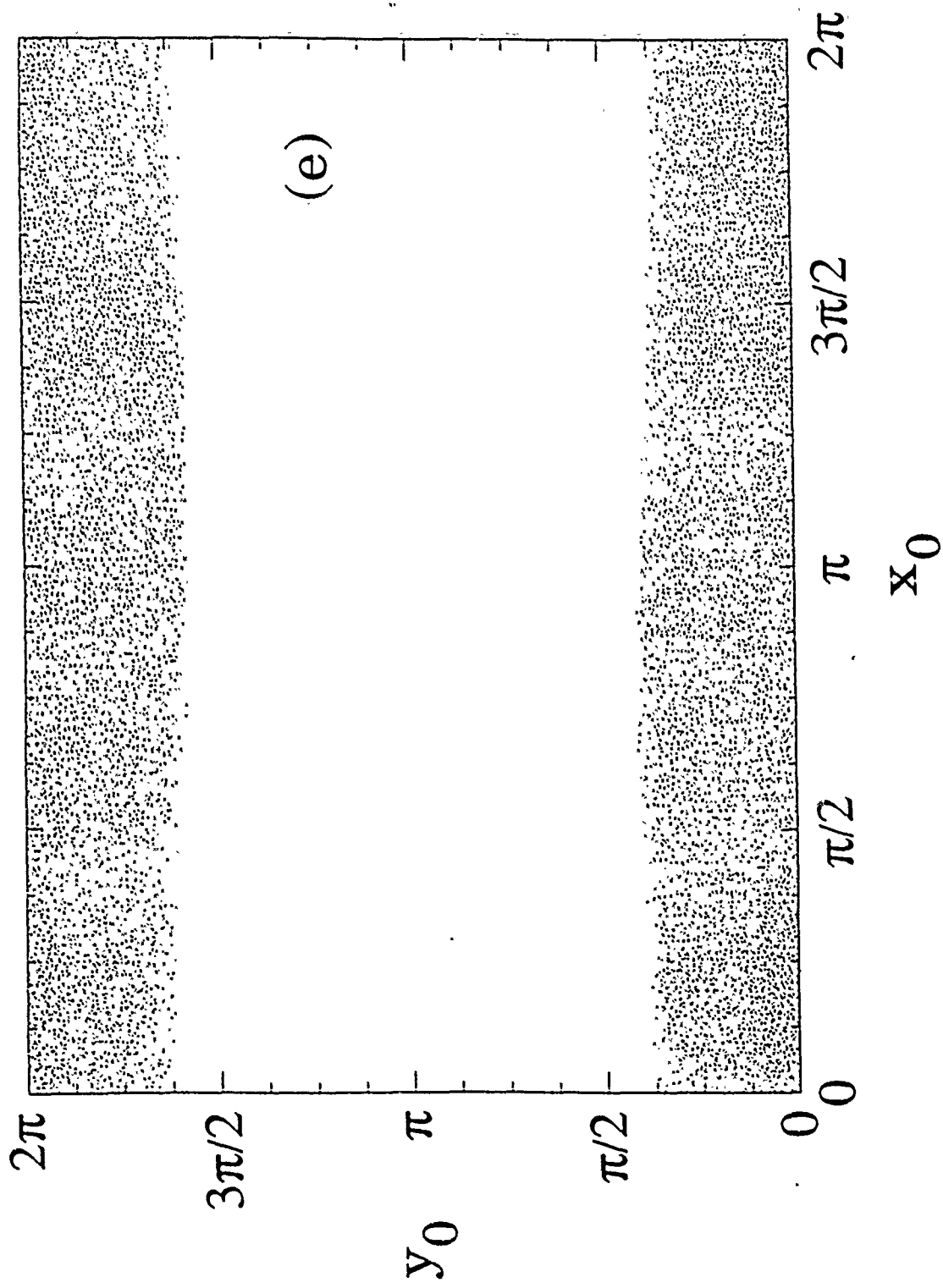


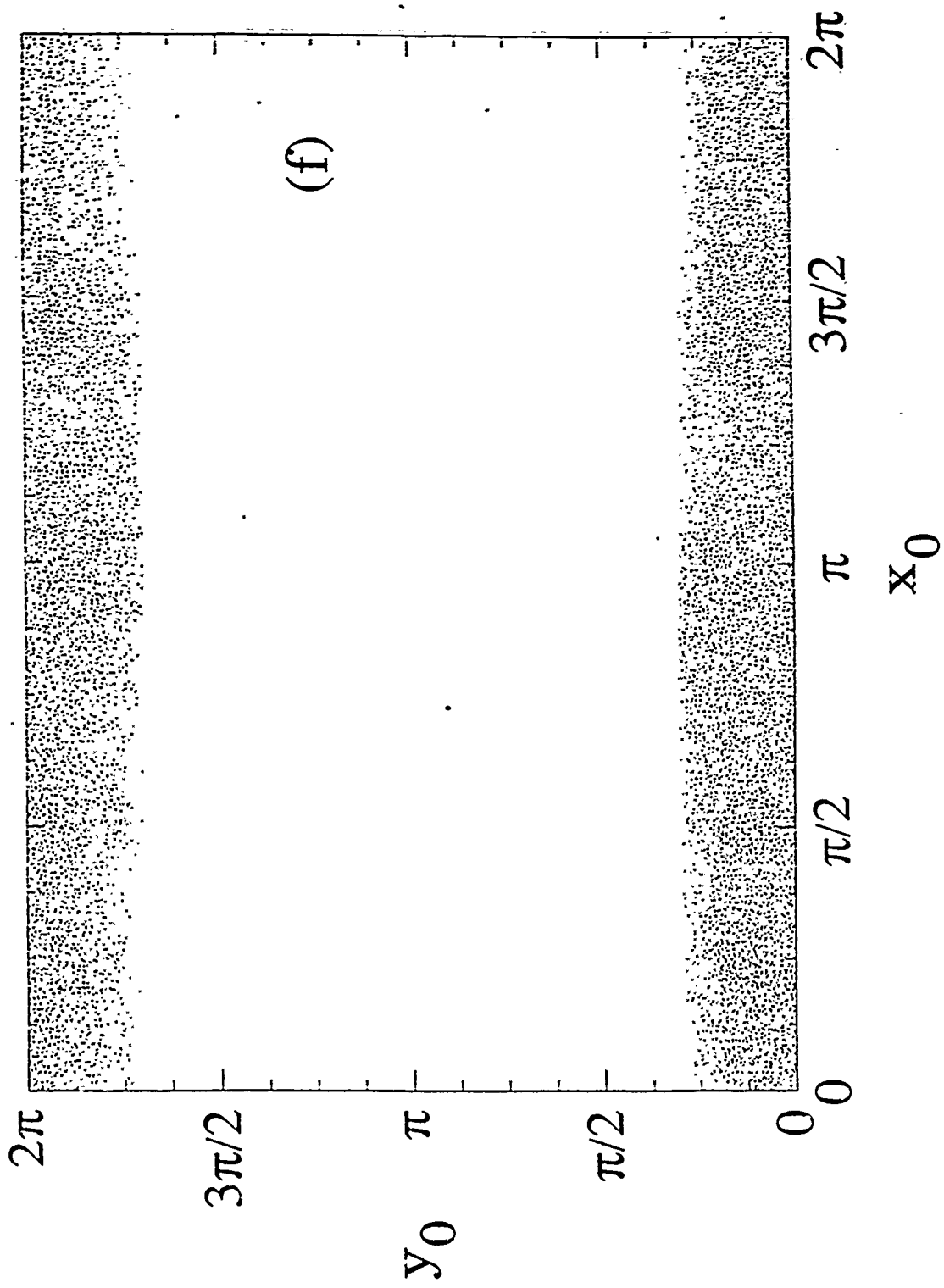


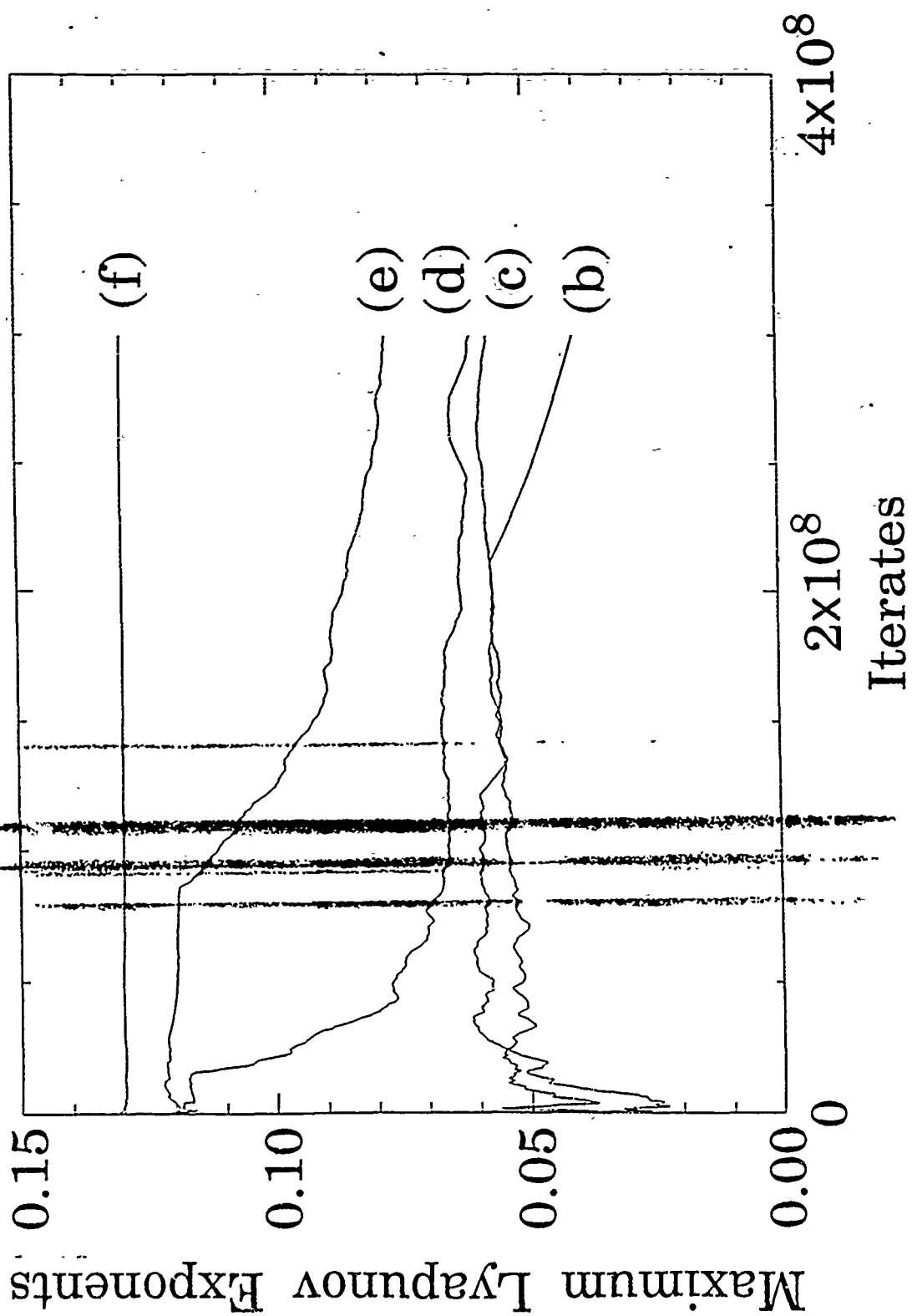


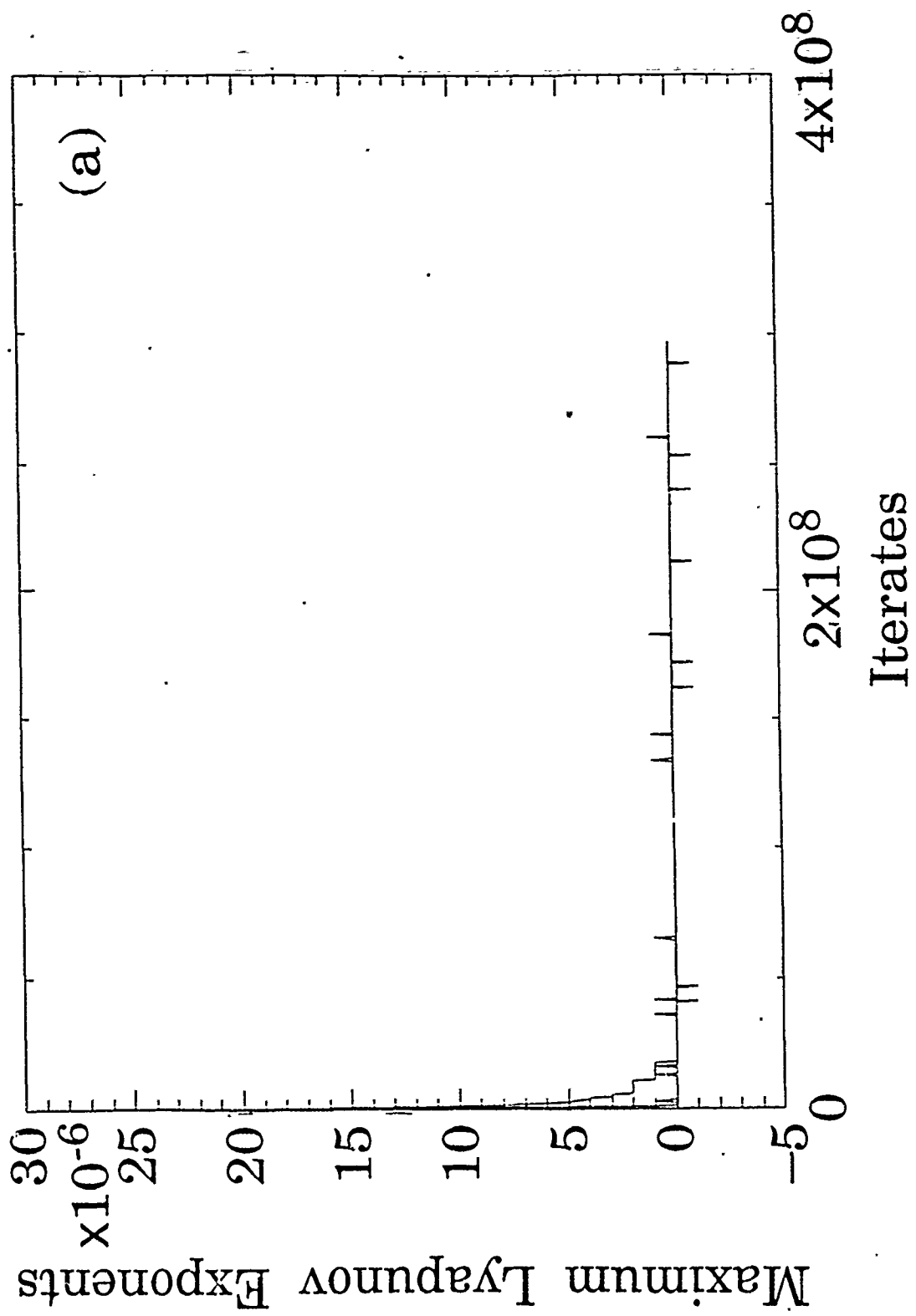


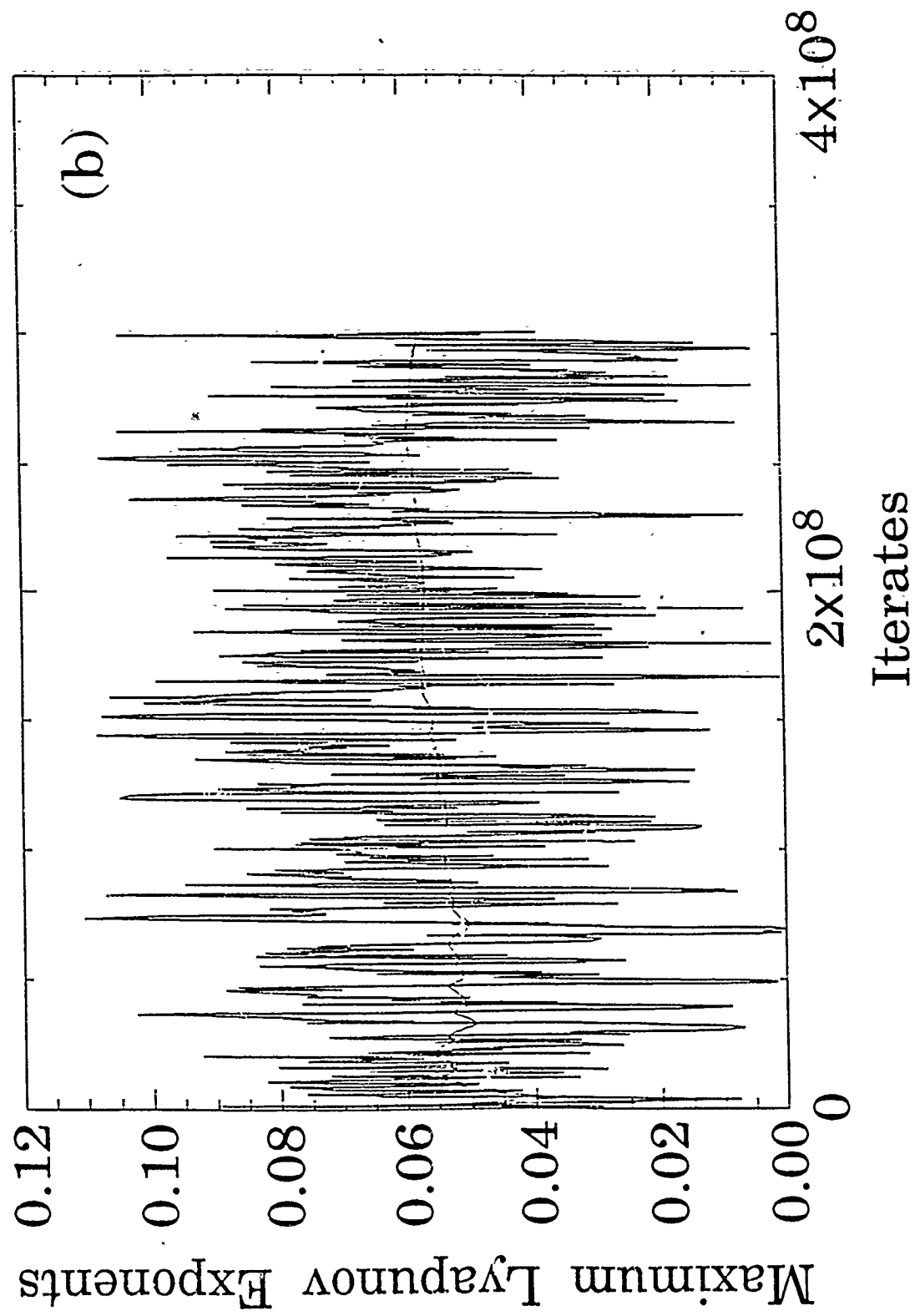


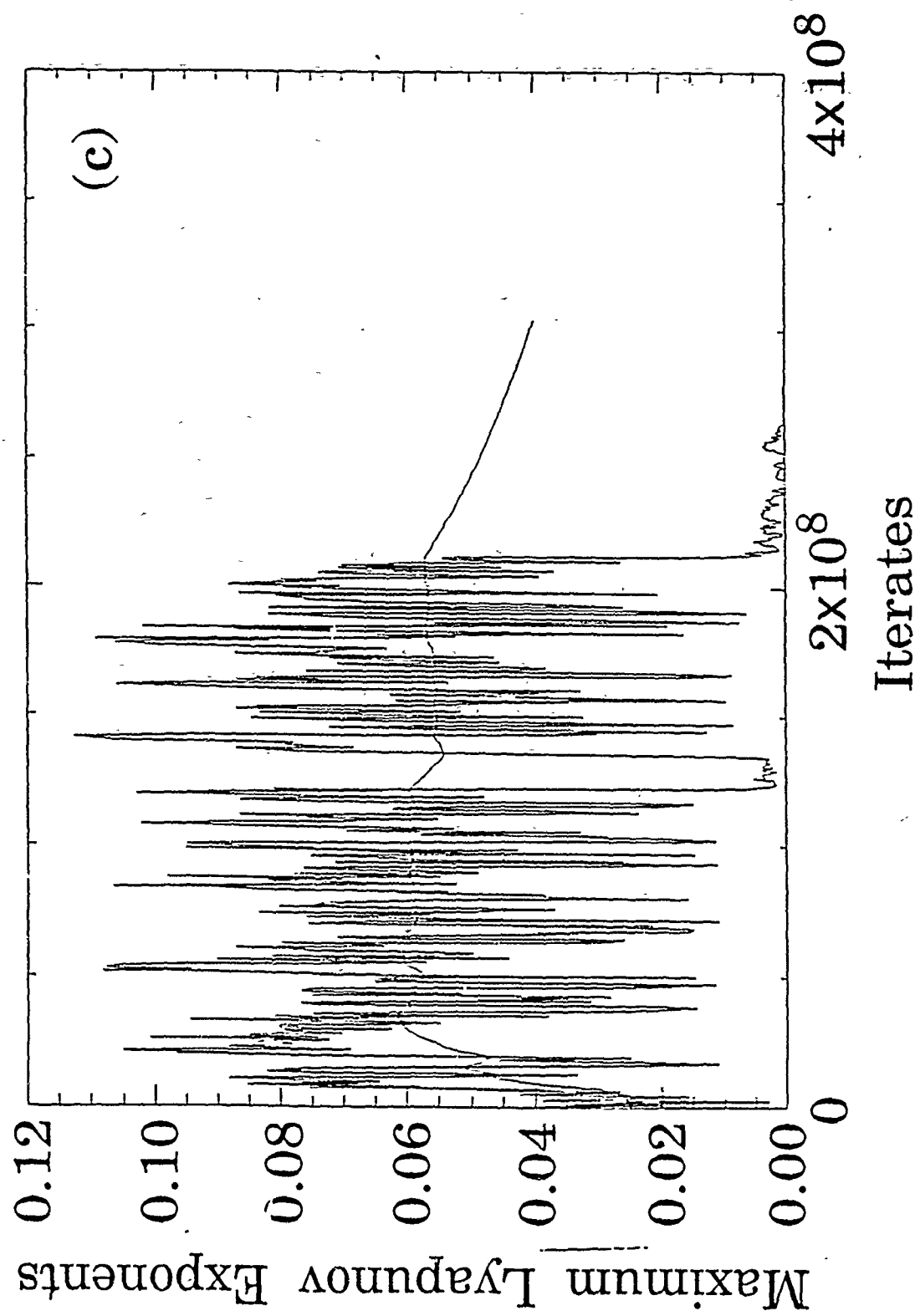


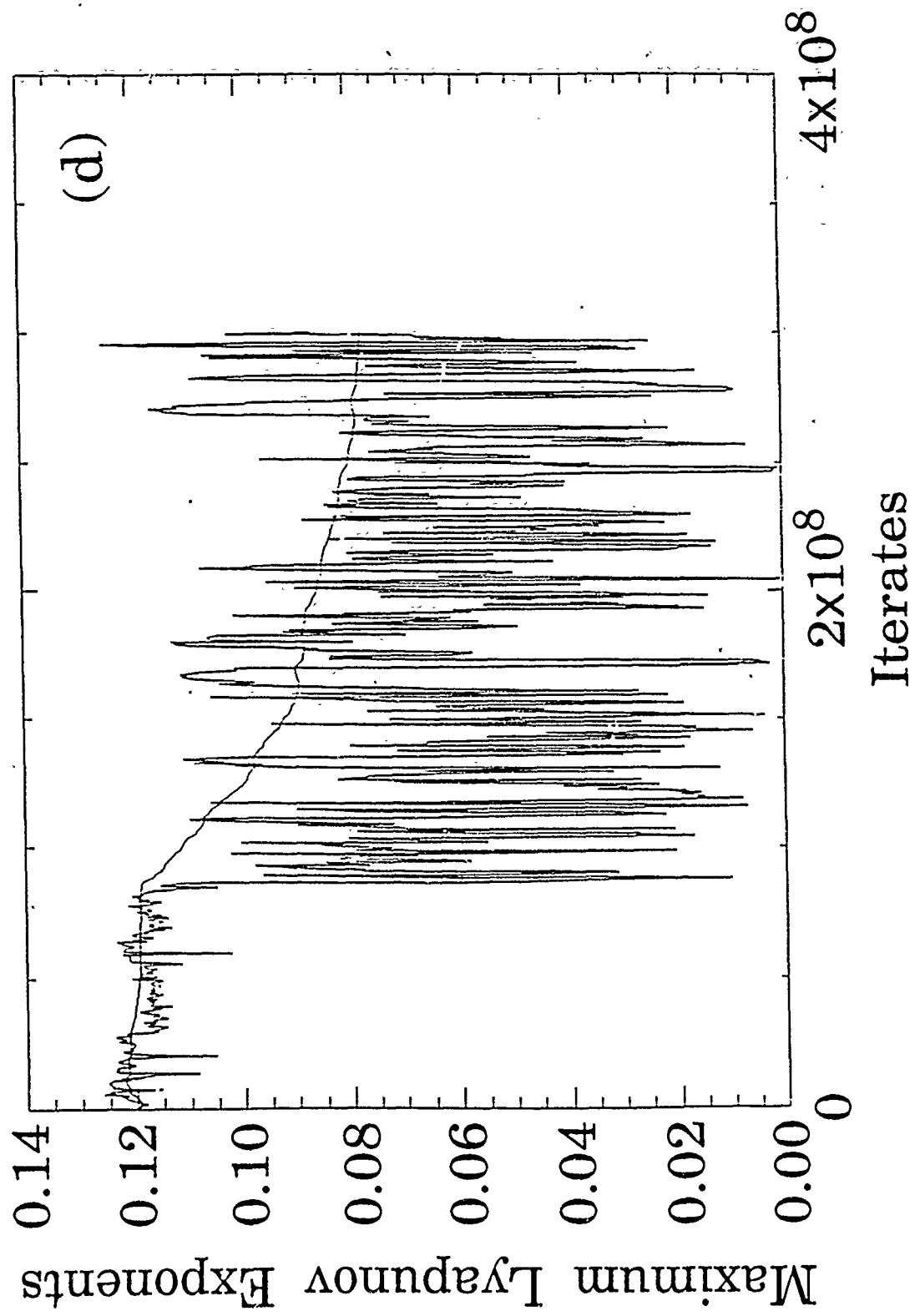


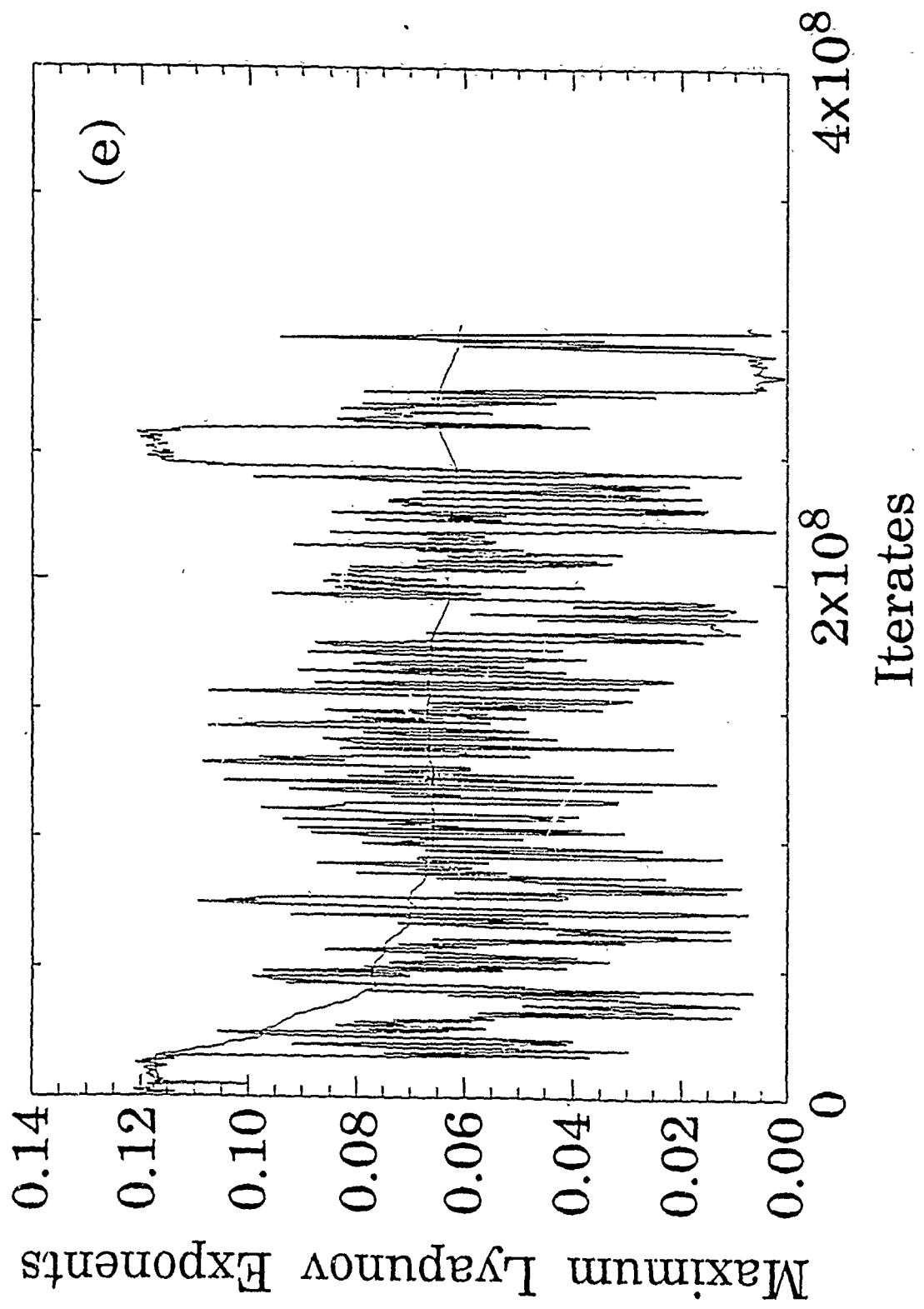


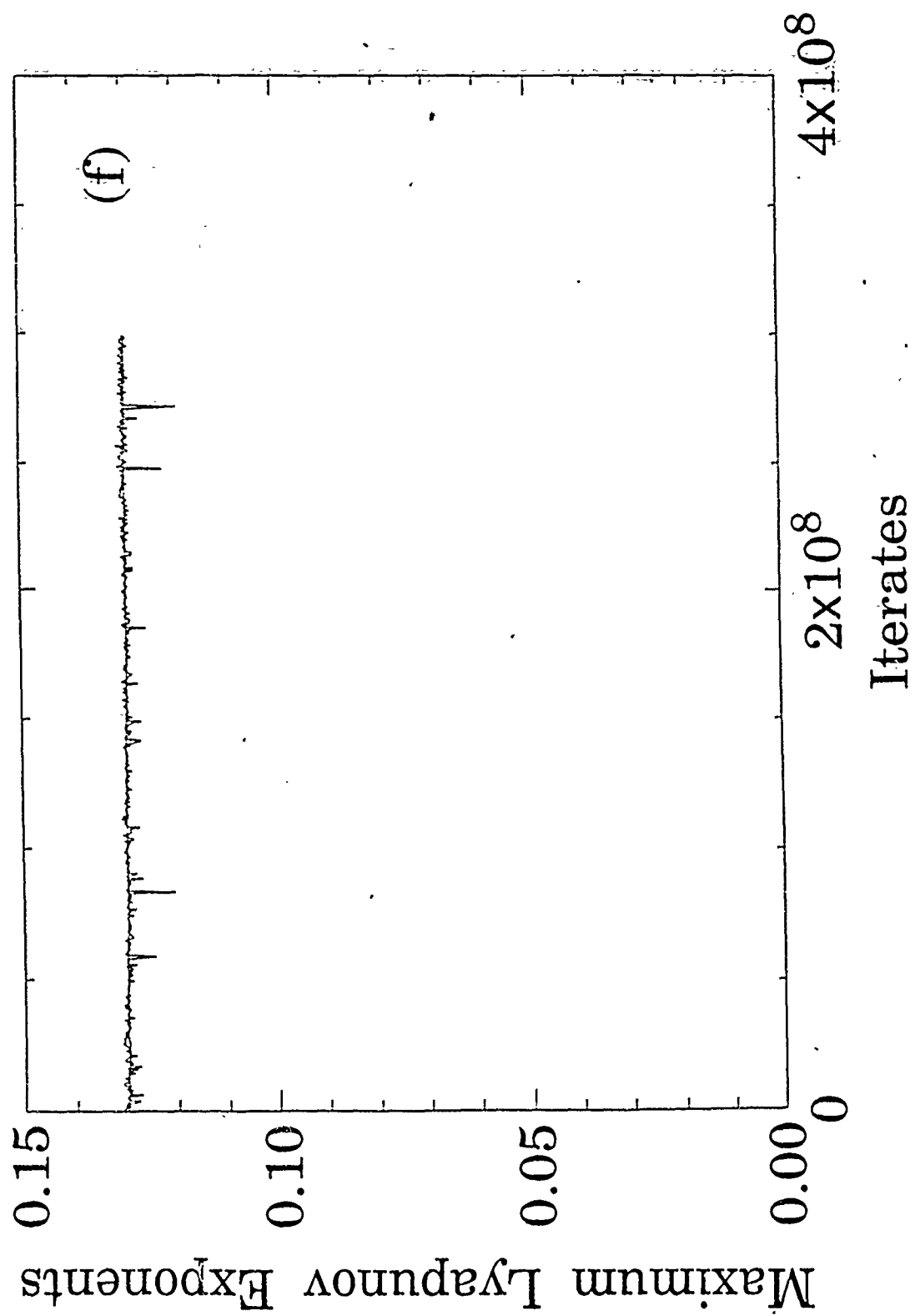












# METAMORPHOSES: SUDDEN JUMPS IN BASIN BOUNDARIES

by

Kathleen T. Alligood  
Department of Mathematics  
George Mason University  
Fairfax, VA 22030

and

Laura Tedeschini-Lalli  
Department of Mathematics and  
Institute for Physical Science and Technology  
University of Maryland, College Park, MD 20742  
On leave from: Dipartimento di Matematica  
Universita di Roma "La Sapienza"  
Rome, Italy I-00185

and

James A. Yorke  
Department of Mathematics and  
Institute for Physical Science and Technology  
University of Maryland  
College Park, MD 20742

December 1985

1) This research was supported in part by grants and contracts from the Defense Advanced Research Projects Agency, The Consiglio Nazionale delle Ricerche (Comitato per le Matematiche), and the Air Force Office of Scientific Research.

## METAMORPHOSES: SUDDEN JUMPS IN BASIN BOUNDARIES<sup>1</sup>

Dynamical systems in the plane can have many coexisting attractors. In order to be able to predict long-term or asymptotic behavior in such systems, it is important to be able to recognize to which attractor (final state) a given trajectory will tend. The set of initial conditions whose trajectories are asymptotic to a particular attractor is called the basin of attraction of that attractor. In some systems that depend on a parameter, it has been observed that the boundaries of these basins are extremely sensitive to small changes in the parameter. Not only can a boundary jump suddenly but it can also change from being smooth to being fractal. These changes, called boundary metamorphoses, are studied at length in [GOY]. In this paper, we prove a theorem, originally stated in [GOV], which characterized the jumps in basin boundaries.

The Hénon map  $f(x,y) = (A-x^2-Jy, x)$  provides an example of this phenomenon. We fix  $J = 0.3$  and vary  $A$ , resulting in a one-parameter, invertible map of the plane. The Jacobian of  $f$  is  $J$ ; hence,  $f$  is area contracting for all  $A$ . We will be looking specifically at the boundary of the basin of attraction of infinity. (The basin of infinity is the set of all points  $(x,y)$  such that  $|f^n(x,y)| \rightarrow \infty$  as  $n \rightarrow \infty$ .) Figures 1a and 1b show the basin of infinity in black for  $A = 1.314$  and  $A = 1.320$ , respectively. In Fig. 1b we see that the basin of infinity contains points which were previously (at  $A = 1.314$ ) well within the white region. This new set of black points has not gradually moved in from the boundary of the white region. Rather, beyond a certain critical value  $A = A^* \approx 1.3145$ , black points suddenly begin appearing deep in the interior of the white region. As  $A$  increases,

Figure 1a.

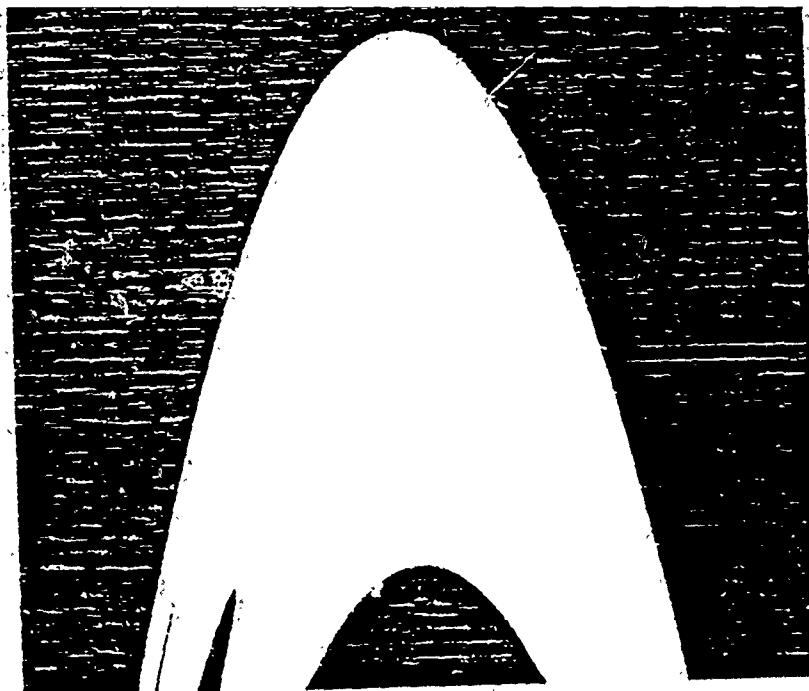


Figure 1b.

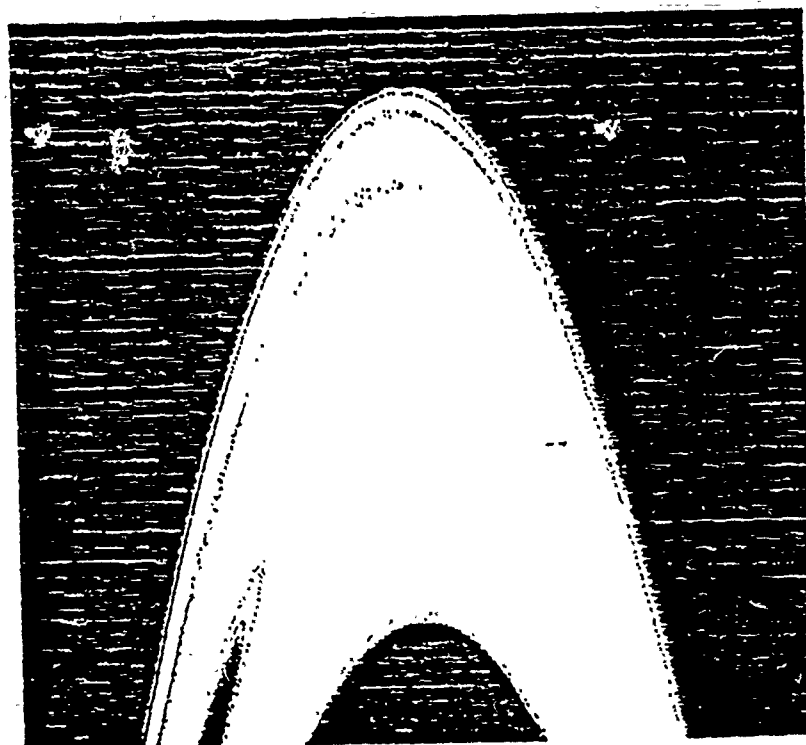


Figure 1 Caption

Figure 1 shows the basin of attraction of infinity in black for the Henon map

$$f(x,y) = (A-x^2-Jy,x).$$

We fix  $J = 0.3$ . In Fig. 1a  $A$  is 1.314, and in Fig. 1b  $A$  is increased to 1.32. The change in the basin of infinity illustrates a basin boundary jump.

the thin bands thicken. This is a discontinuous change in the basin of infinity.

In order to understand this phenomena, we must examine the dynamical behavior on the basin boundary. At  $A = 1.314$  (Fig. 1a) the boundary is observed numerically to consist of a saddle fixed point  $p_1$ , and its stable manifold  $W^S(p_1)$ . (The stable manifold  $W^S(p)$  of a fixed point  $p$  is the set of points  $(x,y)$  such that  $f^n(x,y) \rightarrow p$  as  $n \rightarrow \infty$ . More generally, the stable manifold  $W^S(p_k)$  of a periodic point  $p_k$  of period  $k$  is the set of points  $(x,y)$  such that  $f^{nk}(x,y) \rightarrow p_k$  as  $n \rightarrow \infty$ . Analogously, the unstable manifold  $W^U(p_k)$  of  $p_k$  is the set of points  $(x,y)$  such that  $f^{-nk}(x,y) \rightarrow p_k$  as  $n \rightarrow \infty$ . Such sets can be proved to be smooth curves.) One branch of the unstable manifold of  $p_1$  at  $A = 1.314$  extends into the white region, as shown in Fig. 2a. At the critical value  $A^* = 1.3145$ , after which the basin boundary jumps into the white region, we find that  $W^S(p_1)$  and  $W^U(p_1)$  are tangent (Fig. 2b). S. Hammel and C. Jones [HJ] were the first to prove a theorem relating the tangency of  $W^S(p_1)$  and  $W^U(p_1)$  (called a homoclinic tangency) to basin metamorphoses. Their methods are different from ours, however. We want to relate these metamorphoses to the saddle periodic orbits which are found near the points of tangency and which we describe below.

The complicated dynamical behavior which occurs at homoclinic tangencies has been studied at length in recent years, especially in the papers of Gavrilov and Silnikov [GS], Newhouse [N], and Robinson [R]. Under certain non-degeneracy assumptions, there are horseshoe maps defined on subsets of the plane near a point  $q_0$  of tangency of  $W^S(p_1)$  and  $W^U(p_1)$ . Figure 3 shows a rectangle  $B_4$  and some of its iterates

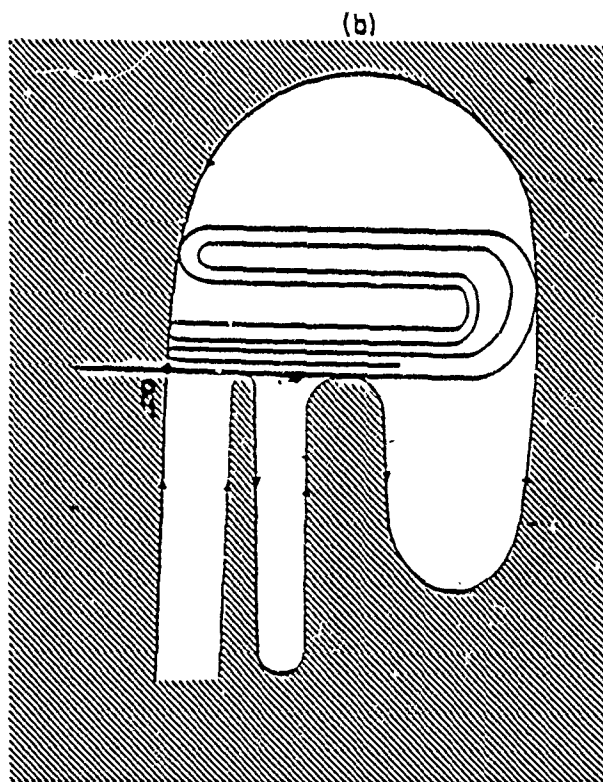
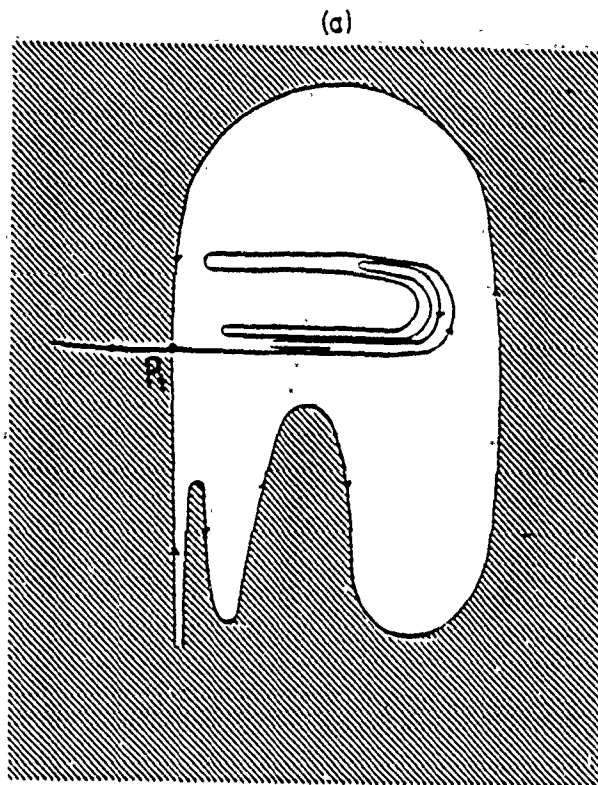


Figure 2 Caption

Figures 2a and 2b show the stable and unstable manifolds of a fixed point  $p_1$  before and at tangency, respectively.

under  $f$ . Notice that  $f^4(B_4)$  is horseshoe shaped and intersects  $B_4$  in two components. In fact, for  $n$  sufficiently large, there is a rectangle  $B_n$  near the point of tangency  $q_0$  such that  $f^n$  restricted to  $B_n$  is a horseshoe map. There is necessarily a saddle orbit of period  $n$  in each of the two components of the intersection of  $B_n$  and  $f^n(B_n)$  (see, for example, [R]). One of these saddles will have a "flipped" unstable manifold (i.e.,  $D_x f^n$  at this saddle has an eigenvalue less than  $-1$ ), and the other will not. We label the unflipped saddle  $p_n$ . This orbit is called a "simple Newhouse periodic orbit" in [TY].

The larger  $n$  is, the closer  $B_n$  will be to  $q_0$  and  $W^S(p_1)$ . This corresponds to the fact that the length of time (i.e., the number of iterates of  $f$ ) it takes for a point to move around the fixed point  $p_1$  is determined by how close the point is to the stable manifold  $W^S(p_1)$ . What we see (Fig. 4) is an infinite family of horseshoes, and a sequence  $\{p_n\}$  of simple Newhouse saddles (where  $p_n$  has period  $n$  and is in  $B_n$ ) such that  $\{p_n\} \rightarrow q_0$ . In the following theorem, as stated in [GOY], the saddle fixed point  $S$  corresponds to  $p_1$  in the discussion above, and the saddle orbit  $T$  corresponds to a simple Newhouse orbit  $p_n$ , for some  $n$ . The term "first non-degenerate tangency" refers to the following set (H) of hypotheses:

- (i)  $W^u(p_1)$  does not intersect  $W^S(p_1)$  for  $A < A^*$ .
- (ii) There exist points  $p_0$  in  $W^u(p_1)$  and  $q_0$  in  $W^S(p_1)$  such that  $q_0 = f^k(p_0)$  for some  $k \geq 1$ , at  $A = A^*$ .
- (iii) There is a parametrization  $h_t$ ,  $-1 \leq t \leq 1$ , of  $W^S(p_1)$  near  $q_0$  such that  $h_0 = q_0$  and  $W^u(p_1)$  near  $q_0$  is given by  $g(h_t)$ , where  $g'(h_0) = 0$  and  $g''(h_0) \neq 0$ .

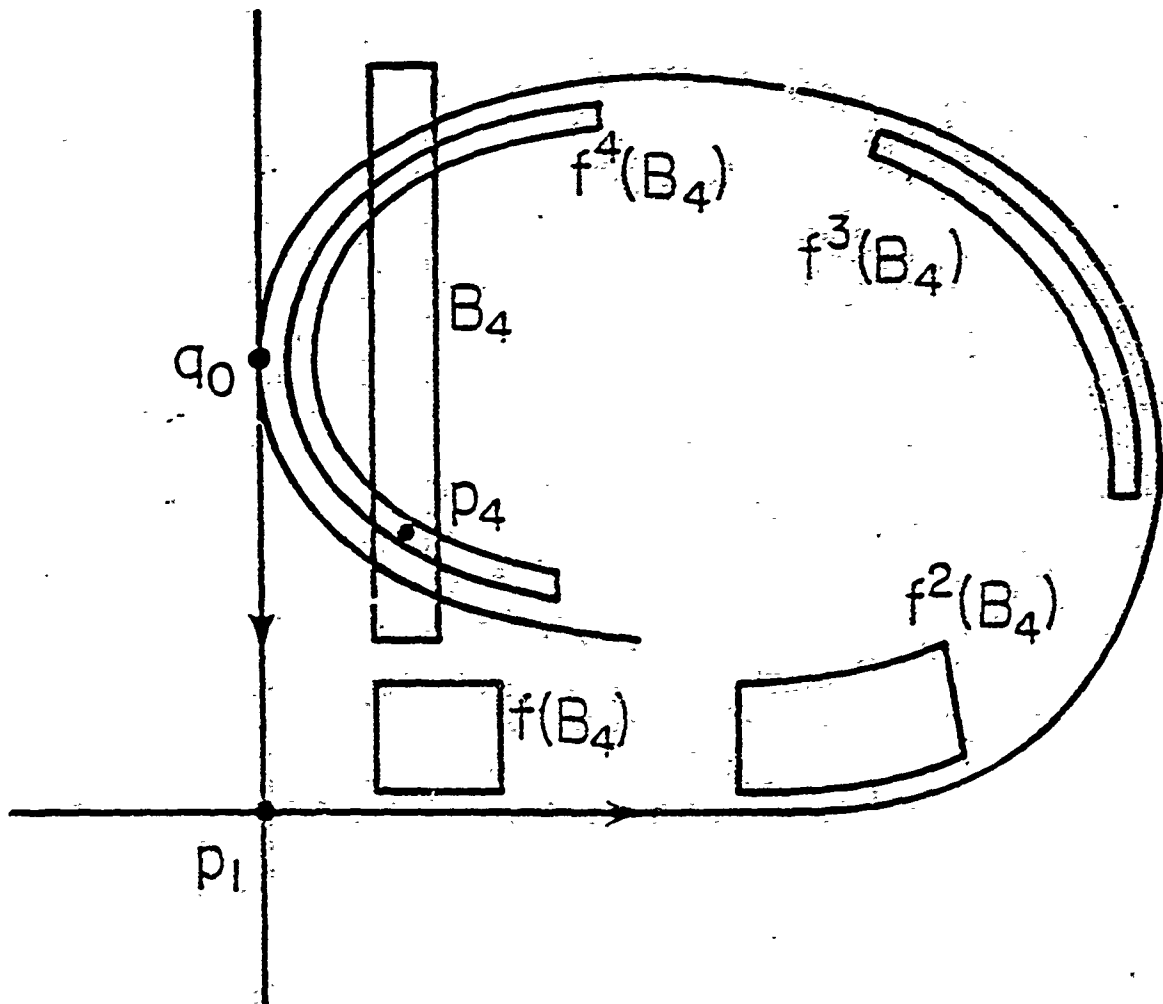


Figure 3 Caption

Figure 3 illustrates a horseshoe map. The invariant set of the horseshoe is in  $B_4 \cap f^4(B_4)$ .

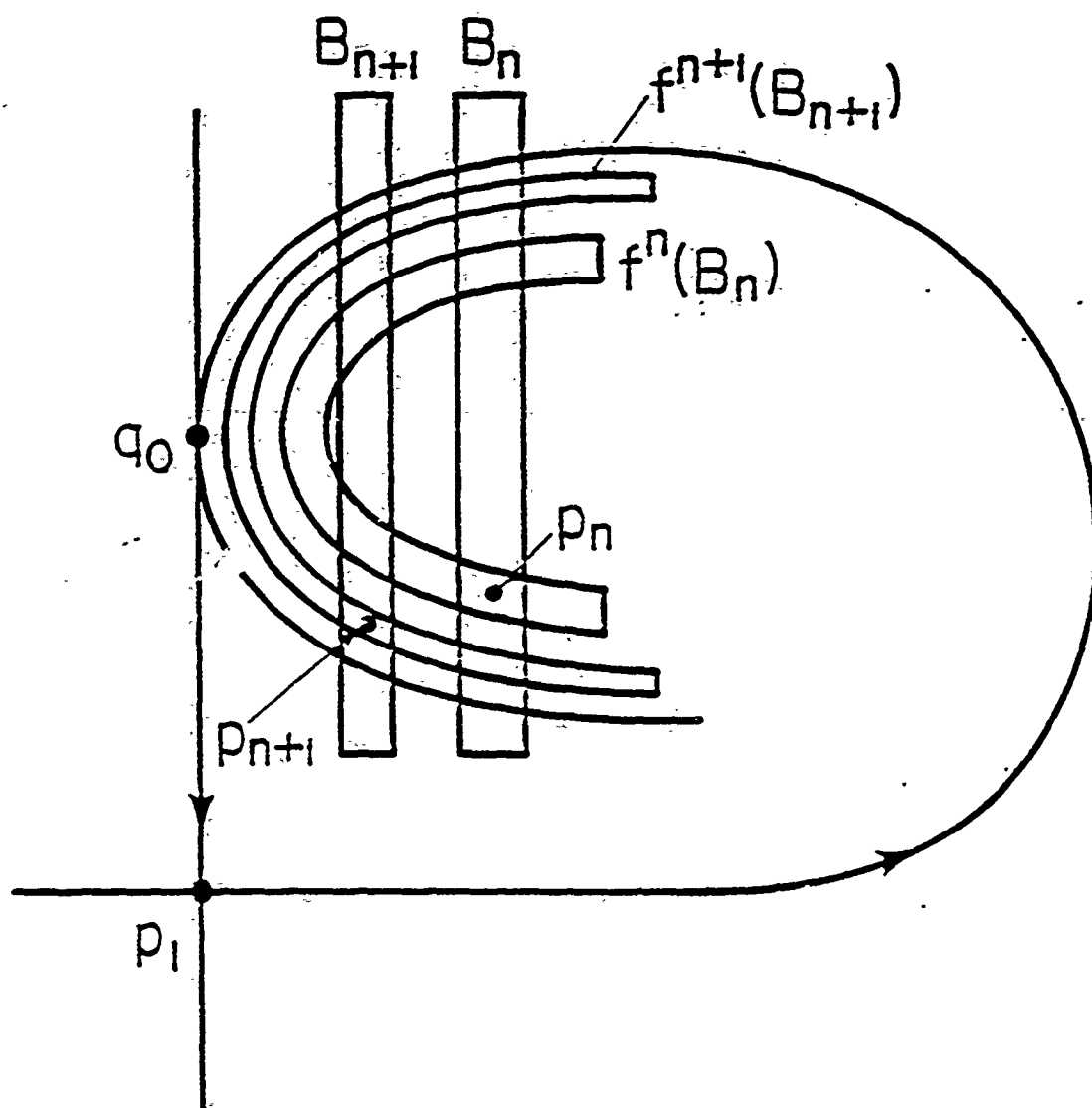


Figure 4 Caption

Figure 4 shows the relative positions of two simple Newhouse saddles  $p_n$  and  $p_{n+1}$  of periods  $n$  and  $n+1$ , respectively.

Theorem. Consider an invertible map  $f$  of the plane depending on a parameter  $A$  with a saddle fixed point or periodic orbit  $S$ . We assume that the absolute value of the determinant of the Jacobian of  $f$  (or of  $f^n$  in the case of a periodic orbit of period  $n$ ) is less than one at every point of the plane. Assume that  $f$  has a transition value  $A^*$  as  $A$  increases where the stable and unstable manifolds of  $S$  have a non-degenerate tangency and then cross for the first time. Then there will be a periodic saddle  $T$  that is in the closure of the stable manifold of  $S$  for all  $A$  slightly greater than  $A^*$  but is not in it at  $A^*$ .

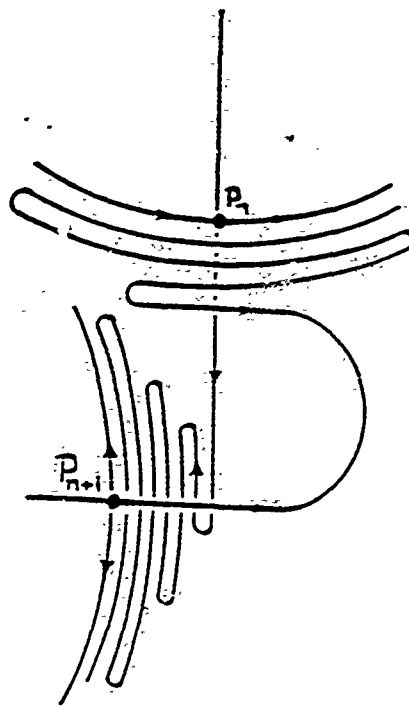
We prove the theorem with the aid of the following lemma.

Lemma. Let  $p_k$  be a simple Newhouse saddle of period  $k$  (as described above) near the point  $q_0$  of tangency of the stable and unstable manifolds of  $p_1$ . Then, for  $n$  sufficiently large, the unstable manifold of  $p_n$  crosses (i.e., intersects transversally) the stable manifold of  $p_{n+1}$ .

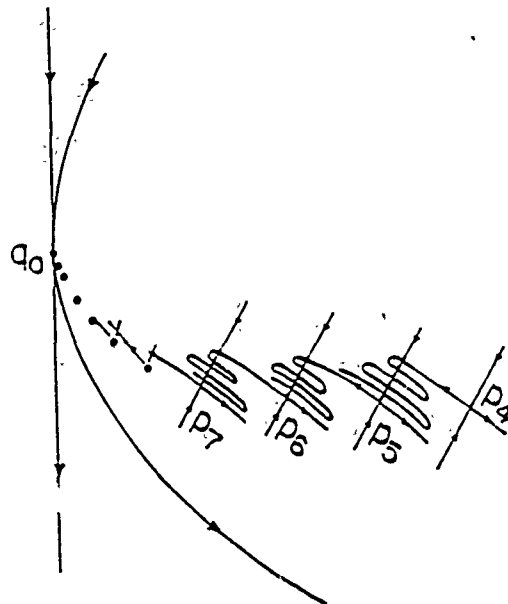
We postpone the proof of this lemma due to its technical nature and proceed to show how the theorem follows. If  $W^u(p_n)$  crosses  $W^s(p_{n+1})$  at a point  $x$ , then the forward iterates of any segment of  $W^u(p_n)$  containing  $x$  will eventually contain all of  $\overline{W^u(p_{n+1})}$  (the closure of  $W^u(p_{n+1})$ ) in its set of limit points<sup>1</sup> (see Fig. 5a). Hence  $\overline{W^u(p_{n+1})} \subset \overline{W^u(p_n)}$ . Proceeding inductively, we have that

---

<sup>1</sup>This follows from the  $\lambda$ -lemma. See, for example, the exposition in [GH].



(a)



(b)

Figure 5 Caption

Figure 5a indicates that the closure of  $W^u(p_{n+1})$  is contained in the closure of  $W^u(p_n)$ . Figure 5b indicates that the point of tangency  $q_0$  is in the closure of the unstable manifolds of infinity many simple Newhouse saddles.

(i)  $\overline{W^u(p_m)} \subset \overline{W^u(p_n)}$ , (see Fig. 5b), for every  $m \geq n$ . In fact, as the proof of the lemma will show, the horseshoe  $f^n(B_n)$  contains a segment of  $W^u(p_n)$  around  $p_n$ . As  $n \rightarrow \infty$ , the horseshoes become thinner and approach  $W^u(p_1)$ . For  $A$  slightly larger than  $A^*$  and for  $m$  sufficiently large, the horseshoe  $F^m(B_m)$  and hence  $W^u(B_m)$  will cut across  $W^s(p_1)$ . thus<sup>1</sup>

$$(ii) \quad \overline{W^u(p_1)} \subset \overline{W^u(p_m)}.$$

Putting together (i) and (ii), we have

(iii)  $\overline{W^u(p_1)} \subset \overline{W^u(p_n)}$ , for  $n$  sufficiently large. Notice that although both  $n$  and  $m$  are taken "sufficiently large" for this argument, the crossing of  $W^u(p_n)$  and  $W^s(p_{n+1})$  occurs for values of  $n$  much smaller than the values of  $m$  for which  $W^u(p_m)$  crosses  $W^s(p_1)$  after tangency. Expression (iii) is equivalent to

$$\overline{W^s(p_n)} \subset \overline{W^s(p_1)}$$

(see, for example, [GOY]). Hence  $p_n$  is in the closure of  $W^s(p_1)$ , for  $A > A^*$ .

#### Remarks.

(1) At  $A = A^*$ , the portion of the plane bounded by  $W^s(p_1)$  from  $q_0$  to  $p_1$  and  $W^u(p_1)$  from  $p_1$  to  $q_0$  is invariant under  $f$ . The saddles  $p_n$  are in the interior of this region, and hence each one is a positive distance from the boundary  $W^s(p_1)$  of the basin of infinity. For every  $A$  slightly larger than  $A^*$ , the theorem says that there is

---

<sup>1</sup>Again by the  $\lambda$ -lemma.

an  $n$  such that  $p_n$  is in the closure of  $W^S(p_1)$ . Thus there is a jump in the boundary at  $A = A^*$ .

(2) The condition that  $n$  is sufficiently large here refers to the value of  $n$  for which the sequence of crossings of  $W^u(p_n)$  and  $W^S(p_{n+1})$  begins. For the Hénon map with  $J = 0.3$  and  $A^* = 1.314$ ,  $n$  appears to be 4 (see [GOY]). This is supported by computer evidence that for  $A$  slightly greater than 1.3145, the saddle  $p_n$  is on the boundary of the basin of infinity.

(3) Non-degeneracy has not been proved for the tangency of the Hénon map at  $A^* = 1.314$ . However, theoretically, almost every such tangency will be non-degenerate.

(4) The proof of the theorem characterizes the boundary after tangency by showing that there are infinitely many saddles and their stable manifolds contained in  $\overline{W^S(p_1)}$ . The fact that there is a jump in the boundary is, of course, implied by this characterization. The existence of such a jump can be demonstrated by a simpler, topological argument. Any path  $I$  connecting the left and right sides of  $B_n$  (cf. Fig. 4) extends through the horse shoe image  $f^n(B_n)$ . If  $f^n(B_n)$  crosses  $B_{n+1}$  (as shown in Fig. 4), a portion of  $f^n(I)$  connects the left and right side of  $B_{n+1}$ . If, at tangency ( $A = A_*$ ),  $f^r(B_r)$  so crosses  $B_{r+1}$  for all  $r$ ,  $r \geq n$ , then  $\bigcup_{r \geq n} f^r(I)$  contains  $q_0$ . For  $A > A_*$ , some forward iterate of  $I$  will then cross  $W^S(p_1)$ .

Proof of Lemma. Following the construction of [R], [TY] (see also [GH; Sec. 6.6]) we assume the following:

(i)  $DF(p_1)$  has eigenvalues  $\nu$  and  $\lambda$  which satisfy  $0 < \nu < 1$ ,  $\lambda > 1$ , and  $\nu\lambda < 1$ .

(ii) There exists a neighborhood  $U$  of  $p_1$  in which the map

$f$  is linear up to smooth changes of coordinates; i.e.,  $f(x,y) = (\lambda x, \nu y)$  for  $(x,y)$  in  $U$ . (Here we need an additional non-resonance assumption--namely, that  $\nu$  and  $\lambda$  are not integer multiples of each other.)

(iii) There is a non-degenerate tangency of  $W^S(p_1)$  and  $W^U(p_1)$ . Specifically, there exist points  $(p_0, 0)$  in  $W^U(p_1)$  and  $(0, q_0)$  in  $W^S(p_1)$  such that  $f^k(p_0, 0) = (0, q_0)$  and  $W^S(p_1)$  and  $W^U(p_1)$  near  $(0, q_0)$  satisfy (H). Furthermore, there is a rectangular neighborhood  $V = [p_0 - \epsilon, p_0 + \epsilon] \times [0, \delta]$  for some  $\epsilon > 0$  and  $\delta > 0$ , such that

$$f^k(x, y) = (\gamma y + \sigma(x - p_0)^2, q_0 - \beta(x - p_0)),$$

for some positive constants  $\gamma, \sigma, \beta$ , all  $(x, y) \in V$ . (See Fig. 6.)

Now let  $W = [0, \sigma\epsilon^2] \times [q_0 - \beta\epsilon, q_0 + \beta\epsilon]$ . (Notice that  $f^k(p_0 \pm \epsilon, 0) = (\sigma\epsilon^2, q_0 \pm \beta\epsilon)$ .) For  $n$  sufficiently large,  $f^{-n+k}(V)$  stretches across  $W$ . For such  $n$ , let  $B_n = f^{-n+k}(V) \cap W$ . Actually, since  $f^{-n+k}(V)$  may wind around a lot, we let  $B_n$  be the connected component of  $f^{-n+k}(V) \cap W$  which is nearest  $W^S(p_1)$ . Under hypothesis (H), we know (see [GH]), that  $f^n$  restricted to  $B_n$  is a horseshoe map, in the sense of Smale [S]. Specifically, we use the following facts about such maps:

(i)  $B_n$  and  $f^n(B_n)$  intersect in two components,  $W_{1,n}$  and  $W_{2,n}$ . The saddle  $p_n$  is contained in  $W_{1,n}$  and is the only fixed point of  $f^n$  in  $W_{1,n}$ . Furthermore,  $p_n$  is the only point in  $W_{1,n}$  which stays in  $W_{1,n}$  under all forward and backward iterates of  $f^n$ .

(ii) The only points which stay in  $W_{1,n}$  under all forward (respectively, backward) iterates of  $f^n$  are in  $W^S(p_n)$  (resp.,  $W^U(p_n)$ ).

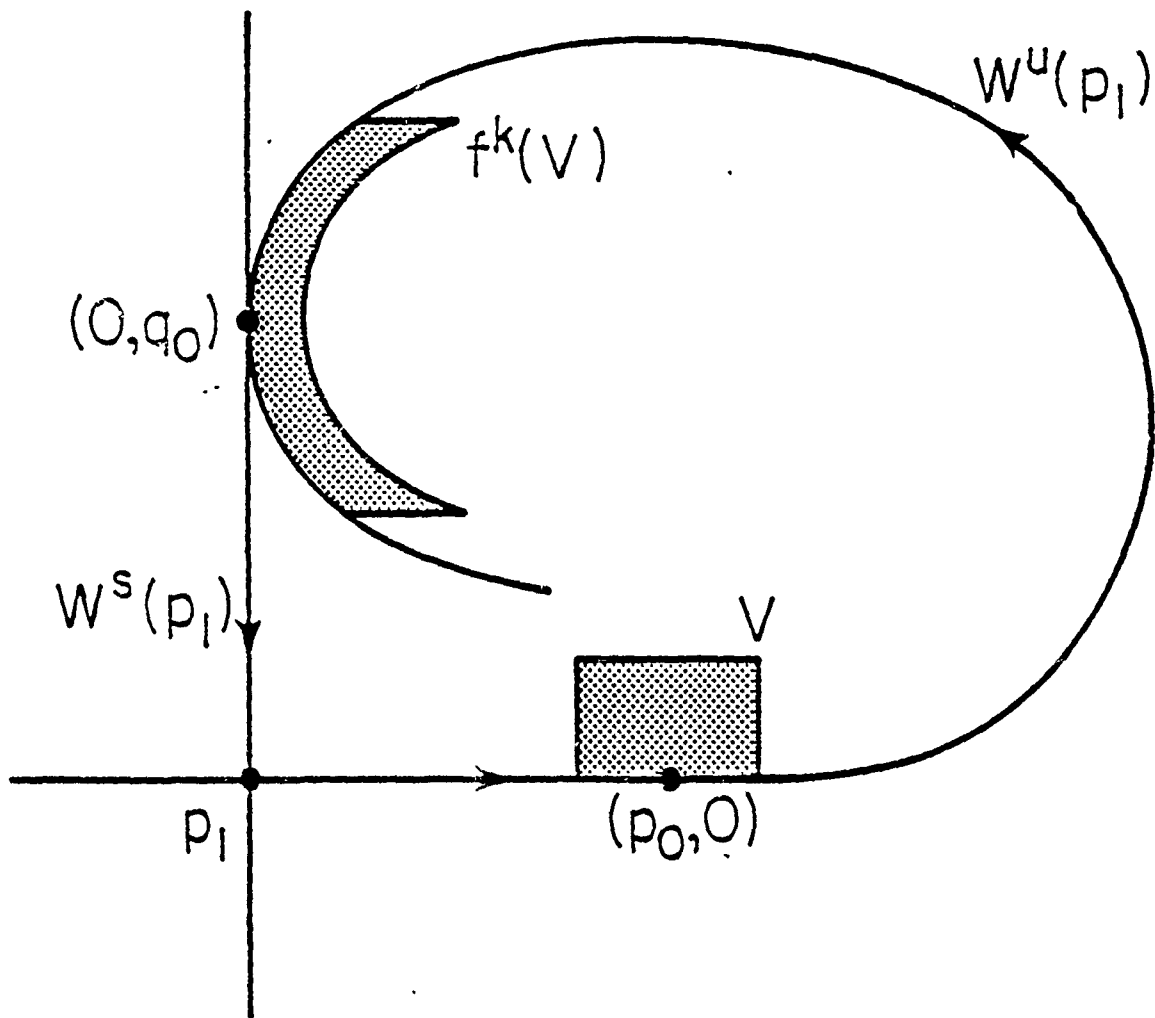


Figure 6 Caption

Figure 6 illustrates definitions used in the proof of the Lemma.

We argue that the stable manifold of  $p_n$  extends (vertically) through  $B_n$  (see Fig. 7). Let  $L_0$  be any horizontal segment in  $B_n$ . It is easily seen that  $f^n(L_0)$  is a parabola which extends through  $f^n(B_n)$ . Recursively, let  $L_i = f^n(L_{i-1}) \cap W_{1,n}$  for  $i = 1, 2, 3, \dots$ . Then  $L_i \subset L_{i-1}$ , and  $\{L_i\}_{i \geq 1}$  is a sequence of nested intervals with  $\text{length}(L_i) < \frac{1}{2} \text{length}(L_{i-1})$ . Hence  $\bigcap_{i \geq 0} L_i$  is one point, call it  $z_0$ . Since  $f^m(z_0)$  is in  $W_{1,n}$  for all  $m > 1$ ,  $z_0$  must be in  $W^s(p_n)$ . This argument shows that  $W^s(p_n)$  intersects the top and bottom of  $B_n$  and first leaves  $B_n$  through these sides. A similar argument (using iterates of  $f^{-1}$ ) shows that  $W^u(p_n)$  extends through the horseshoe  $f^n(B_n)$ , first leaving the horseshoe through the "feet". (See Fig. 7).

In order to prove that  $W^u(p_n)$  intersects  $W^s(p_{n+1})$ , we need to show that the horseshoe  $f^n(B_n)$  containing  $W^u(p_n)$  crosses through  $B_{n+1}$  (see Fig. 8). Let  $Q$  be the distance from  $(0, q_0)$  to  $B_{n+1}$ , and let  $P$  be the distance from  $(0, q_0)$  to the vertex of the right parabola boundary of  $F^n(B_n)$ , as shown in Fig. 8. It is easily seen by our assumptions on  $f$  that  $Q = \lambda^{-(n+1)+k}(p_0 - \epsilon)$  and  $P = \gamma \nu^{n-k}(q_0 + \beta \epsilon)$ . We conclude that  $\frac{P}{Q} = \lambda \gamma \left( \frac{q_0 + \beta \epsilon}{p_0 - \epsilon} \right) (\lambda \rho)^{n-k} \rightarrow 0$  as  $n \rightarrow \infty$ , since  $\lambda \rho < 1$ .  $\square$

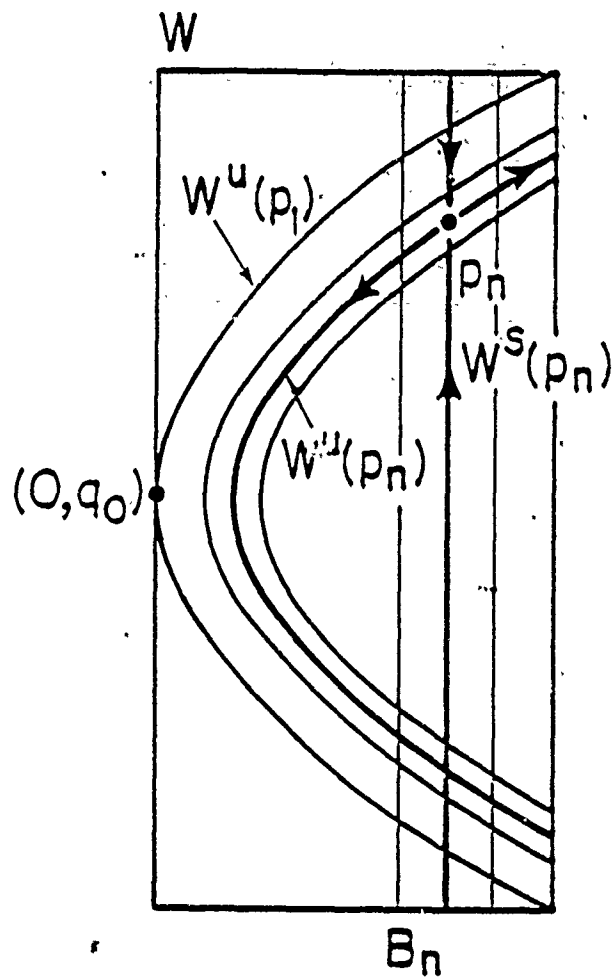


Figure 7 Caption

Figure 7 shows parts of the stable and unstable manifolds of the simple Newhouse saddle  $p_n$ .

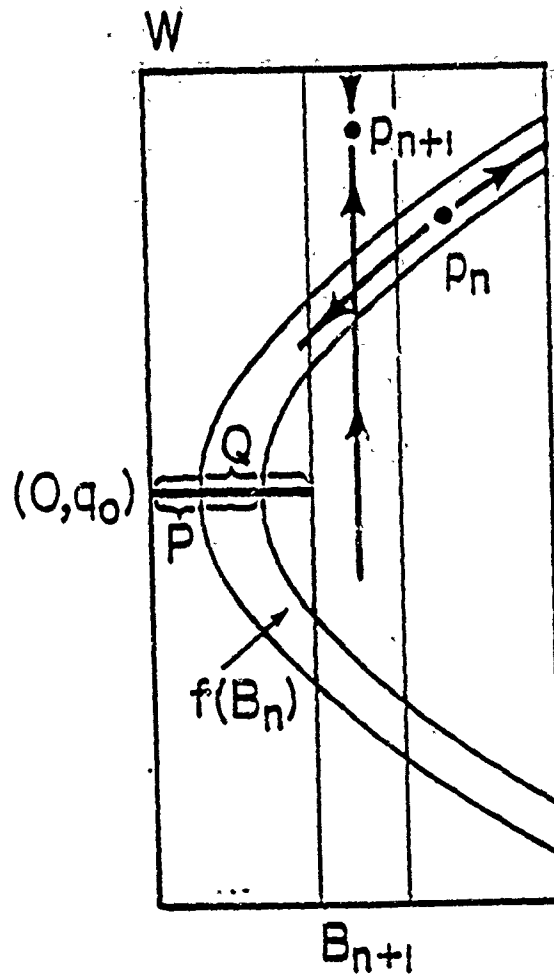


Figure 8 Caption

Figure 8 illustrates definitions used in the proof of the Lemma.

## REFERENCES

- [GH] J. Guckenheimer and P. Holmes, Nonlinear Oscillations, Dynamical Systems, and Bifurcation of Vector Fields, Berlin, Heidelberg, New York: Springer, 1983.
- [GOY] C. Grebogi, E. Ott, and J. Yorke, Phys. Rev. Lett. 56 (1986), 1011.
- [GS] N. Gavrilov and L. Silnikov, "On three-dimensional dynamical systems close to systems with structurally unstable homoclinic curve I," Math. USSR Sbornik 17 (1972), 467-485; and part II, Math. USSR Sbornik 19 (1973), 139-156.
- [HJ] S. Hammel and C. Jones, personal communication; and "A dissipative map of the plane--a model for optical bistability," Doctoral Dissertation by S. Hammel, University of Arizona, 1986.
- [N] S. Newhouse, "Diffeomorphisms with infinitely many sinks," Topology 13 (1974), 9-18.
- [R] C. Robinson, "Bifurcation to infinitely many sinks," Commun. Math. Phys. 90 (1983), 433-459.
- [S] S. Smale, "Differentiable dynamical systems," Bull. Am. Math. Soc. 73 (1967), 747-817.
- [TY] L. Tedeschini-Lalli and J. Yorke, "How often do simple dynamical processes have infinitely many sinks?" preprint

# The Analysis of Experimental Data Using Time-Delay Embedding Methods

Eric J. Kostelich\*

James A. Yorke

Institute for Physical Science and Technology

University of Maryland

College Park, Maryland 20742

January 30, 1989

## Abstract

The time delay embedding method provides a powerful tool for the analysis of experimental data, including a novel method for noise reduction. In addition, we describe how the method allows experimentalists to use many of the same techniques that have been essential for the analysis of nonlinear systems of ordinary differential equations and difference equations.

## 1 Introduction

Numerical computation and computer graphics have been essential tools for investigating the behavior of nonlinear maps and differential equations. The pioneering work of Lorenz [24] was made possible by numerical integration on a computer, allowing him to take nearby pairs of initial conditions and compare the trajectories. Hénon [23] discovered the complex dynamics of his celebrated quadratic map with the aid of a programmable calculator. A

---

\*Mailing address: Center for Nonlinear Dynamics, Department of Physics, University of Texas, Austin, Texas 78712

variety of classical and modern techniques has been exploited to find periodic orbits; their stable and unstable manifolds [20]; basins of attraction [25]; fractal dimension [26]; and Lyapunov exponents [17, 29, 35]. In some cases, numerical methods can establish rigorously the existence of initial conditions whose trajectories have essentially the same intricate structure that one sees on a computer screen [11].

Unfortunately, until now experimentalists have not been able to apply most of these methods to the analysis of experimental data, since they do not in general have explicit equations to model the behavior of their apparatus. In cases where it is possible to find accurate models of the physical system, quantitative predictions about the behavior of actual experiments are possible [22]. However, all that is available in a typical experiment is the time dependent output (e.g. voltage) from one or more probes, which is a function of the dynamics. Until recently, power spectra have been the principal method for analyzing such data. For instance, Fenstermacher *et al.* [19] relied heavily on power spectra to detect transitions from periodic to weakly turbulent flow between concentric rotating cylinders. However, Fourier analysis alone is inadequate for describing the dynamics.

Other methods have been used to analyze time series output from dynamical systems. For instance, Lorenz [24] used next amplitude maps to describe some features of the dynamics; that is, he plotted  $z_{n+1}$  against  $z_n$  where  $z_n$  is the  $n$ th relative maximum of the third coordinate of the numerically calculated solution. Such maps are often useful, not only for investigating features of the Lorenz attractor [30], but also for instance in experiments on intermittency in oscillating chemical reactions [28].

In the past several years, the so-called *embedding method* has come into common use as a way of reconstructing an attractor from a time series of experimental data. In this approach, one supposes that the dynamical behavior is governed by a solution traveling along an attractor<sup>1</sup> (which is not observable directly). However, one assumes that there is a smooth function which maps points on the attractor to real numbers (which are the experimental measurements). In the embedding method, one generates a set of  $m$ -dimensional points whose coordinates are values in the time series separated by a constant delay [9]. For example, when  $m = 3$ , the reconstructed attractor is the set of points  $\{x_i = (s_i, s_{i+\tau}, s_{i+2\tau})\}$  where  $\tau$  is the time delay.

---

<sup>1</sup>Existing numerical methods require the attractor to be low dimensional.

Takens [32] has shown that under suitable hypotheses, this procedure yields a set of points which is equivalent to points on the original attractor.

The earliest applications of the embedding method may be called *static* in that the analysis focuses on the geometric properties of the set of points on the reconstructed attractor. For example, phase portraits and Poincaré sections are used in [4] to help determine the transition between quasiperiodic and chaotic flow in a Couette-Taylor experiment. Another important static method is the estimation of attractor dimension from experimental data, for which there is a large literature [26]. In addition, various information theoretic notions can be used to find good choices of embedding dimension and time delay [21].

Certain recent applications of the embedding method are quite different in nature and can be called *dynamic* in that information about the dynamics is stored in the computer for analysis. With each data vector  $x_i$ , one stores the "next" vector, for example,  $x_{i+\delta}$  for some  $\delta > 0$ . This makes it possible to compute a linear approximation of the dynamics in a neighborhood of  $x_i$ , assuming that there is a low dimensional dynamical system underlying the data.<sup>2</sup> In particular, a linear approximation provides an estimate of the Jacobian of the map at  $x_i$  [9]. Eckmann and Ruelle [17] use linear maps computed in this way to integrate a set of variational equations and find the positive Lyapunov exponents.<sup>3</sup>

In fact, the embedding method provides a powerful set of tools for analyzing the dynamics, the breadth of which may not have been realized by Eckmann and Ruelle. In this paper, we discuss two novel applications that are possible, specifically:

- **Noise reduction.** Since one can approximate the dynamics at each point, it becomes possible to identify and correct inaccuracies in trajectories arising from errors in the original time series. Numerical evidence suggests that the noise reduction procedure described below improves the accuracy of other analyses, such as Lyapunov exponents and dimension calculations.
- **Simplicial approximations.** Linear approximations can be computed at each point on a grid in a neighborhood of the attractor to

---

<sup>2</sup>This material was first presented by D. Ruelle at a Nobel symposium in June 1984?

<sup>3</sup>Wolf et al. [35] have proposed a different method in which nearby pairs of points are followed to estimate the largest Lyapunov exponent.

form a simplicial approximation of the dynamical system. This can be used to locate unstable periodic orbits near the attractor.

We begin with a description of noise reduction in the next section.

## 2 Noise Reduction

The ability to extract information from time varying signals is limited by the presence of noise. Recent experiments to study the transition to turbulence in systems far from equilibrium, like those by Fenstermacher *et al.* [19], Behringer and Ahlers [1], and Libchaber *et al.* [16], succeeded largely because of instrumentation that enabled them to quantify and reduce the noise. However, it is often expensive and time consuming to redesign experimental apparatus to improve the signal to noise ratio.

In cases where the time series can be viewed as a dynamical system with a low dimensional attractor, the time delay embedding method can be exploited to correct errors in trajectories that result from noise. This is done in two steps once an embedding dimension  $m$  and a time delay  $\tau$  have been fixed. In the first step, we consider the motion of an *ensemble* of points in a small neighborhood of each point on the attractor in order to compute a linear approximation of the dynamics there. In the second step, we use these approximations to consider how well an *individual* trajectory obeys them. That is, we ask how the observed trajectory can be perturbed slightly to yield a new trajectory that satisfies the linear maps better. The trajectory adjustment is done in such a way that a new time series is output whose dynamics are more consistent with those on the phase space attractor.

This approach is fundamentally different from traditional noise reduction methods. Because we consider the motion of points on a phase space attractor, we are using information in the original signal that is not localized in a time or frequency domain. Points which are close in phase space correspond to data which in general are widely and irregularly spaced in time, due to the sensitive dependence on initial conditions on chaotic attractors. In contrast, Kalman [3] and similar filters examine data which are closely spaced in time; Wiener [27] filters operate in the frequency domain.

### 3 Eckmann-Ruelle linearization

The discrete sampling of the original signal means that the points on the reconstructed attractor can be treated as iterates of a nonlinear map  $f$  whose exact form is unknown. We assume that  $f$  is nearly linear in a small neighborhood of each attractor point  $\mathbf{x}$  and write

$$f(\mathbf{x}) \approx A\mathbf{x} + \mathbf{b} \equiv L(\mathbf{x})$$

for some  $m \times m$  matrix  $A$  and  $m$ -vector  $\mathbf{b}$ . (The matrix  $A$  is the Jacobian of  $f$  at  $\mathbf{x}$ .)

This approximation, which we call the *Eckmann-Ruelle linearization at  $\mathbf{x}$* , can be computed with least squares methods similar to those described in [9, 17]. Given a reference point  $\mathbf{x}_{\text{ref}}$ , let  $\{\mathbf{x}_i\}_{i=1}^n$  be a collection of the  $n$  points which are closest to  $\mathbf{x}_{\text{ref}}$ . With each point  $\mathbf{x}_i$  we store the next point (i.e., the image of  $\mathbf{x}_i$ ), denoted  $\mathbf{y}_i$ .<sup>4</sup> The  $k$ th row  $\mathbf{a}_k$  of  $A$  and the  $k$ th component  $b_k$  of  $\mathbf{b}$  are given by the least squares solution of the equation

$$y_k = b_k + \mathbf{a}_k \cdot \mathbf{x}, \quad (1)$$

where  $y_k$  is the  $k$ th component of  $\mathbf{y}$  and the dot denotes the dot product. Figure 1 illustrates the idea.<sup>5</sup>

We mention three difficulties in computing the local linear approximations in the subsections below.

#### 3.1 Ill conditioned least squares

There is a particular problem when one tries to compute solutions to Eq. 1 with a finite data set of limited accuracy that has not been addressed in previous papers [17, 29]. Suppose for example that all the points in a neighborhood of  $\mathbf{x}_{\text{ref}}$  lie nearly along a single line, i.e., the attractor appears one dimensional within the available resolution. Although it is possible to measure the expansion along the unstable manifold at  $\mathbf{x}_{\text{ref}}$ , there are not enough

<sup>4</sup>The points  $\mathbf{x}_i$  are points on the attractor which are *not* consecutive in time. The subscript  $i$  merely enumerates all the points on the attractor contained within a small distance  $\epsilon$  of  $\mathbf{x}_{\text{ref}}$ . In this notation,  $\mathbf{x}_i$  and  $\mathbf{y}_i$  are consecutive in time.

<sup>5</sup>Farmer and Sidorowich [18] observe that the Eckmann-Ruelle linearization can be used for prediction. Given a reference point  $\mathbf{x}_i$ , find the Eckmann-Ruelle linearization  $A_i\mathbf{x} + \mathbf{b}_i$ , compute  $\mathbf{x}_{i+1} = A_i\mathbf{x}_i + \mathbf{b}_i$ , and repeat the process to get the predicted trajectory.

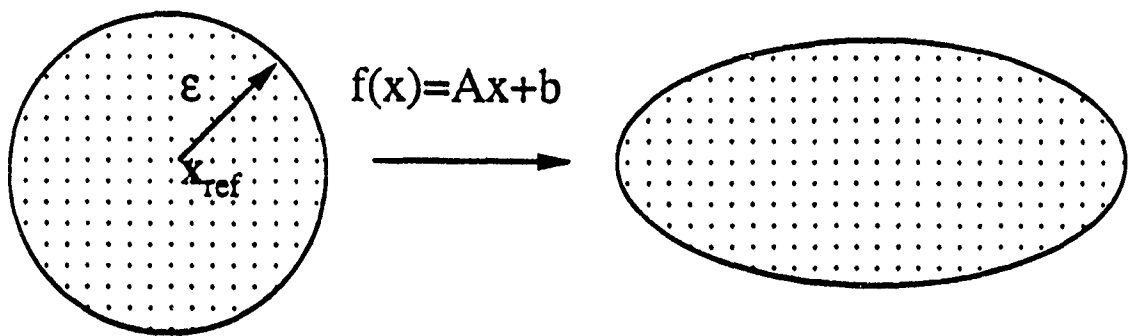


Figure 1: Schematic diagram for the first stage of the noise reduction method. A collection of points in an  $\epsilon$ -ball about the reference point  $\mathbf{x}_{\text{ref}}$  is used to find a linear approximation of the dynamics there.

points in other directions to measure the contraction. Hence it is not possible to compute a  $2 \times 2$  Jacobian matrix accurately. Any attempt to do so will result in an estimate of the Jacobian whose elements have large relative errors. This kind of least squares problem is *ill conditioned*.

The ill conditioning can be avoided by changing coordinates so that the first vector in the new basis points in the unstable direction.<sup>6</sup> A one dimensional approximation of the dynamics is computed using the new coordinates; that is, we approximate the dynamics only along the unstable manifold. We recover the matrix  $A$  by changing coordinates back to the original basis.

For example, if we are working in the plane and the unstable direction is the line  $y = x$ , then we rotate the coordinate axes by 45 degrees. The dynamics are approximated by a one-dimensional linear map computed along the line  $y = x$ . Then we rotate back to the original coordinates. (The resulting matrix  $A$  has rank 1 in this example.) This approach substantially enhances the robustness of the numerical procedure.

### 3.2 Finding nearest neighbors

A second problem is finding an efficient way to locate all of the points closest to a given reference point. The dynamical embedding method imposes stringent requirements on any nearest-neighbor algorithm. The storage overhead for the corresponding data structures must be small, because there are tens of thousands of attractor points. The algorithm must be fast, since there is one nearest-neighbor problem for each linear map to be computed.

We solve this problem by partitioning the phase space into a grid of boxes that is parallel to the coordinate axes. Each coordinate axis is divided into  $B$  intervals. (Figure 2 illustrates the grid in two dimensions.) Each point on the attractor is assigned a box number according to its coordinates. For example, a point on the plane whose first coordinate falls in the  $j$ th interval (counting from 0) along the  $x$  axis and whose second coordinate falls in the  $k$ th interval along the  $y$  axis is assigned to box number  $kB + j$ . The list of box numbers is sorted, carrying along a pointer to the original data point. Given a reference point  $x_{\text{ref}}$ , its box number is found using the above formula. A binary search in the list of box numbers then locates the address of  $x_{\text{ref}}$

---

<sup>6</sup>This is done by computing the right singular vectors [8] of the  $n \times m$  matrix whose  $j$ th row is  $x_j$ .

$B^2 - B$	$B^2 - B + 1$	$B^2 - B + 2$	$\dots$	$B^2 - 1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$B$	$B + 1$	$B + 2$	$\dots$	$2B - 1$
0	1	2	$\dots$	$B - 1$

Figure 2: Box numbering scheme in 2 dimensions. The attractor is normalized to fit in the unit square. The bottom row of boxes rests against the  $x$  axis and the leftmost row of boxes against the  $y$  axis.

and all the other points in the same box number. The search is extended if necessary to adjacent boxes.

Only a crude partition is needed for this algorithm to work efficiently (typically we choose  $B = 40$ ), and the grid is extended only to the first three coordinate axes. When the embedding dimension is larger than three, a preliminary list of nearest neighbors is obtained using only the first three coordinates of each attractor point. The final list is extracted by computing the distances from  $\mathbf{x}_{ref}$  to each point in the preliminary list.

Although there are circumstances where this algorithm can perform poorly (e.g., when most of the attractor points are concentrated in a handful of boxes), the distribution of points on typical attractors is sufficiently uniform that the running time is very fast. Memory use is also efficient: a set of  $N$  attractor points requires  $3N$  storage locations. In contrast, the tree-search algorithm advocated in [18] requires several times more storage (although the lookup time is probably slightly less). Because  $N \approx 10^5$  in typical applications, we believe that the box-grid approach (or some variant) is the most practical. A survey of other nearest-neighbor algorithms is given in [2].

### 3.3 Errors in variables

There is a potential difficulty in the use of ordinary least squares to compute the linear maps. In the usual statistical problem of fitting a straight line, one has observations  $(x_i, y_i)$  where  $x_i$  is known exactly and  $y_i$  is measured. One assumes that  $y_i = a_0 + a_1 x_i + \epsilon_i$ , where the  $\epsilon_i$  are independent errors drawn from the same normal distribution. (Analogous assumptions hold

in the multivariate case.) In the present situation, however, both  $x_i$  and  $y_i$  are measured with error. It can be shown that ordinary least squares produces biased estimates of the parameters  $a_0$  and  $a_1$  in this case [15, 10]. In practice this does not seem to be a serious problem, but statistical procedures to handle this situation (the so-called "errors in variables" methods) may provide an alternative approach to noise reduction. We consider this question in the appendix.

## 4 Trajectory Adjustment by Minimizing Self Inconsistency

The Eckmann-Ruelle linearization procedure described above is computed and the resulting maps are stored for a sequence of reference points along a given trajectory (for the results quoted here, the sequence usually contains 24 points). We now consider how to perturb this trajectory so that it is more consistent with the dynamics. The objective is to choose a new sequence of points  $\hat{x}_i$  to minimize the sum of squares

$$\sum w \|\hat{x}_i - x_i\|^2 + \|\hat{x}_i - L_{i-1}(\hat{x}_{i-1})\|^2 + \|\hat{x}_{i+1} - L_i(\hat{x}_i)\|^2 \quad (2)$$

where  $L(x_i) = A_i x_i + b_i$ ,  $w$  is a weighting factor, and the sum runs over all the points along the trajectory.<sup>7</sup> Equation 2 can be solved using least squares. Heuristically, Eq. 2 measures the self-inconsistency of the data, assuming that the linear approximations of the dynamics are accurate. See Fig. 3. We say the new sequence  $\{\hat{x}_i\}$  is more *self consistent*.

The trajectory adjustment can be iterated. That is, once a new trajectory  $\hat{x}_i$  has been found, one can replace each  $x_i$  in Eq. 2 by  $\hat{x}_i$  and compute a new sequence  $\{\hat{x}_i\}$ .

We place an upper limit on the distance a point can move. Points which seem to require especially large adjustments can be flagged and output unchanged. (This may be necessary if the input time series contains large

<sup>7</sup>In the results described in this paper, the Eckmann-Ruelle linearization procedure is done using a collection of points within a radius of 1-6% of the each reference point, depending on the embedding dimension, the dimension of the attractor, and the number of attractor points. This results in collections of 50-200 points per ball, which gives reasonably accurate map approximations without making the computer program too slow. The weighting factor  $w$  is set to 1.

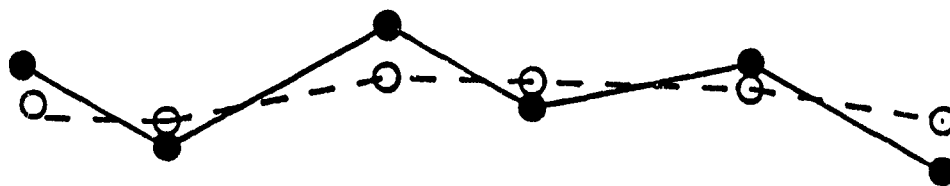


Figure 3: Schematic diagram of the trajectory adjustment procedure. The trajectory defined by the sequence  $\{x_i\}$  is perturbed to a new trajectory given by  $\{\hat{x}_i\}$  which is more consistent with the dynamics. In this example the dashed line shows what the perturbed trajectory might look like if the dynamics were approximately horizontal translation to the right.

“glitches” or if nonlinearities are significant over small distances in certain regions of the attractor.)

When the input is a time series, we modify the above procedure slightly since we require a time series as output. The trajectory adjustment is done so that changes to the coordinates of  $x_i$  (corresponding to particular time series values) are made consistently for all subsequent points whose coordinates are the same time series values. For example, suppose the time delay is 1 and the embedding dimension is 2. Then trajectories are perturbed so that the second coordinate of the  $i$ th point is the same as the first coordinate of the  $(i + 1)$ st point. That is, when  $x_i = (s_i, s_{i+1})$  is moved to the point  $\hat{x}_i = (\hat{s}_i, \hat{s}_{i+1})$ , we require that the first coordinate of  $\hat{x}_{i+1}$  be  $\hat{s}_{i+1}$ .

## 5 Results using experimental data

We note that the attractor need not be chaotic for this noise reduction procedure to be effective. Fig. 4(a) shows a phase portrait of noisy measurements of wavy vortex flow in a Couette-Taylor experiment [12]. This flow is periodic, so the attractor is a limit cycle (widened into a band because of the noise) and the power spectrum consists of one fundamental frequency and its harmonics above a noise floor. See Fig. 4(b). Figures 4(c)-(d) show the same data after noise reduction. The noise reduction procedure makes the limit cycle much narrower, and the noise floor in the power spectrum is reduced by almost two orders of magnitude. However, no power is subtracted from any of the fundamental frequencies, and in fact some harmonics are revealed which previously were obscured by the noise.

These results are significantly different from those obtained by low pass filtering. Figure 4(e)-(f) shows the phase portrait and power spectrum when the original data are passed through a 12th-order Butterworth filter with a cutoff frequency of 0.35. The dynamical noise reduction procedure is more effective than low pass filtering since the noise appears to have a broad spectrum.

However, the method appears to subtract power from a mode whose fundamental frequency is approximately 0.3 times the Nyquist frequency. We do not know exactly why this occurs. However, this peak corresponds to the rotation frequency of the inner cylinder and may result from a defect in the Couette-Taylor apparatus [31]. We do not consider this to be a serious problem, because the power associated with this mode is several orders of magnitude smaller than that of the wavy vortex flow.

We emphasize that our objective is to find a simple dynamical system that is consistent with the data. It is possible for this method to eliminate certain dynamical behavior from an attractor if those dynamics have small amplitude. This situation is most likely to arise when there are not enough data to distinguish such dynamics from random noise. In the present example, the noise reduction procedure reveals the limit cycle behavior quite well.<sup>8</sup>

The results obtained by applying the method to chaotic data from the

---

<sup>8</sup>We have not attempted to find the smallest amplitude at which the noise reduction procedure can distinguish quasiperiodic from periodic flow.

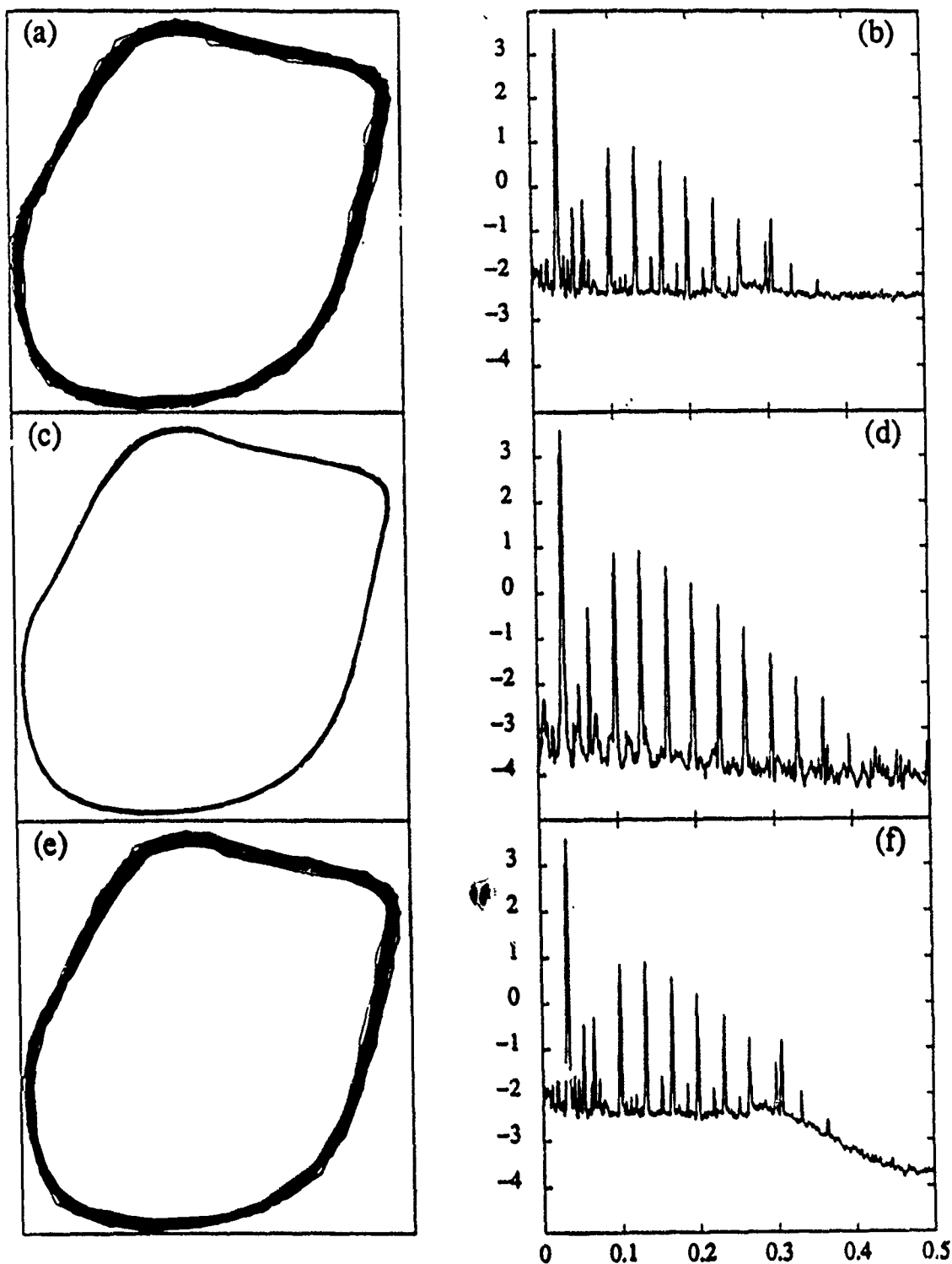


Figure 4: Phase portraits and power spectra for measurements of wavy vortex flow in a Couette-Taylor experiment. (a)-(b) Phase portrait and power spectrum before noise reduction is applied; (c)-(d) after noise reduction; (e)-(f) after a low pass filter is applied to the original data. The vertical axis in (b), (d) and (f) is the base-10 logarithm of the power spectral density; the horizontal axis is in multiples of the Nyquist frequency.

Couette-Taylor fluid flow experiment described in [4] are shown in Fig. 5. Figure 5(a) shows a two dimensional phase portrait of the raw time series at a Reynolds number  $R/R_c = 12.9$ , which corresponds to weakly chaotic flow [4]. The corresponding phase portrait from the filtered time series is shown in Fig. 5(b). Figs. 5(c)-(d) show the power spectra for the corresponding time series.<sup>9</sup>

It is difficult to estimate how much noise is removed from the data in this example on the basis of power spectra. One problem is that the transition from quasiperiodic to weakly chaotic fluid flow is marked by a sudden rise in the noise floor in the power spectrum (cf. Fig. 3 in [4]). Hence one cannot determine how much of the noise floor is due to deterministic chaos and how much results from broadband noise. The noise reduction procedure described here has the effect of reducing the power in the high frequency components of the signal. One question therefore is whether reducing the high-frequency noise corresponds to discovering the true dynamics which have been masked by noise. We believe that the answer is yes, based on those cases where there is an underlying low-dimensional dynamical system. However, in chaotic processes some high-frequency components remain, because they are appropriate to the dynamics.

## 6 Numerical Experiments on Noise Reduction

One important question is how much noise this method removes from the data. The power spectra above suggest that the method eliminates most of the noise, but it is impossible to give a precise estimate for typical experimental data.

However, the Hénon map [23] provides a convenient way to quantify the noise reduction, because it can be written as a time delay map of the form

$$x_{i+1} = f(x_i, x_{i-1}) = 1 - \alpha x_i^2 + \beta x_{i-1}. \quad (3)$$

We use Eq. 3 to generate a time series as follows (with the standard parameter values  $\alpha = 1.4$ ,  $\beta = 0.3$ ). We choose an initial condition and discard the

---

<sup>9</sup>The time series consists of 32,768 values, from which an attractor is reconstructed in four dimensions. Linear maps are computed using 50-100 points in each ball. Trajectories are fitted using sequences of 24 points.

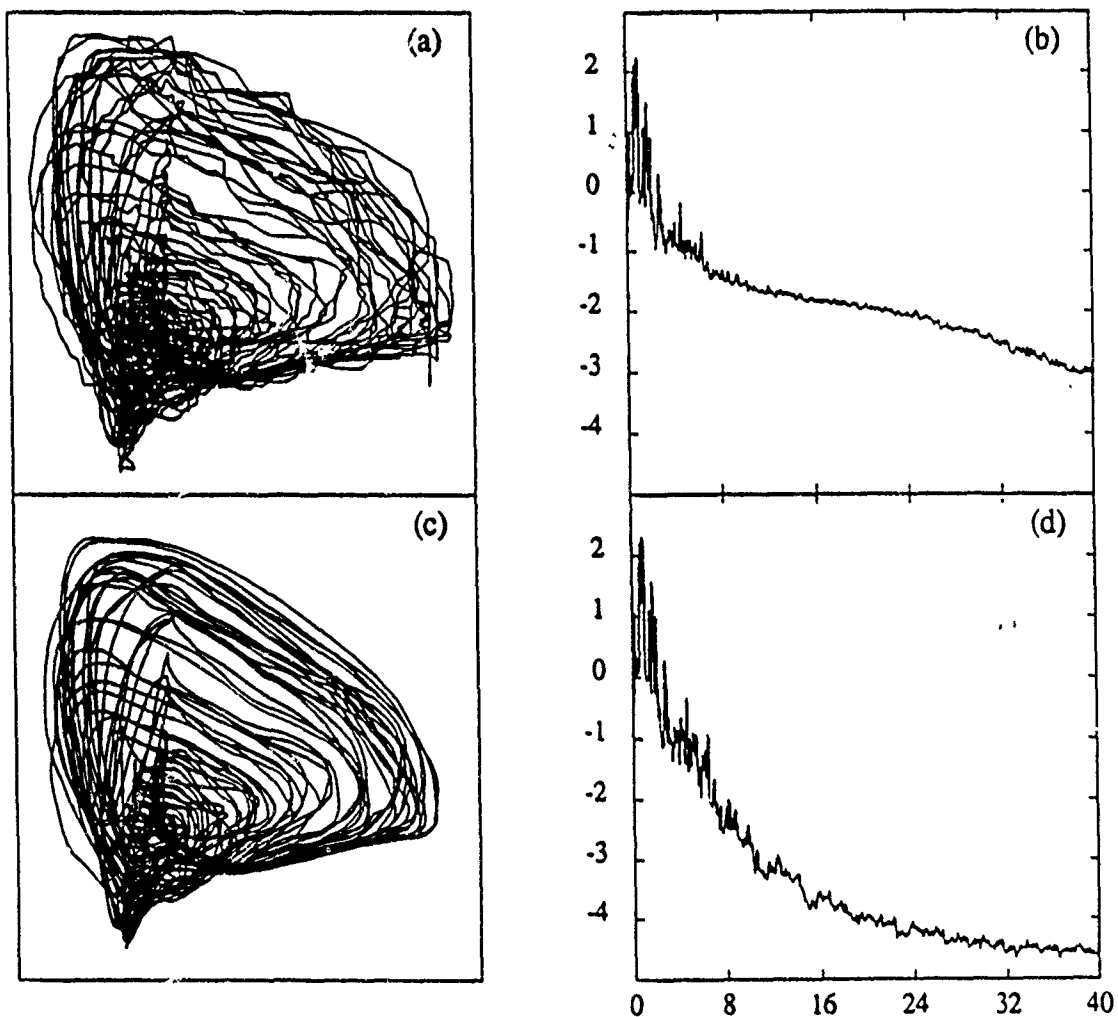


Figure 5: Phase portraits and power spectra for measurements of weakly chaotic flow in a Couette-Taylor experiment. (a)-(b) Phase portrait and power spectrum before noise reduction is applied; (c)-(d) after noise reduction. The units for the power spectrum plots are the same as those in [4].

first 100 iterates. The next 32,768 iterates are stored, and a time series is generated by adding a uniformly distributed random number to each iterate. This simulates a time series with *measurement noise*, i.e., a time series where noise results from errors in measuring the signal, not from perturbations of the dynamics.

We measure the improvement in the signal after processing by considering the *pointwise error*  $e_i = \|x_{i+1} - f(x_i, x_{i-1})\|$ , i.e., the distance between the observed image and the predicted one. Let the *mean error* be  $E = (\sum e_i^2/N)^{1/2}$ , the rms value of the pointwise error over all  $N$  points on the attractor. We define the *noise reduction* as  $R = 1 - E_{\text{fitted}}/E_{\text{noisy}}$ , where the mean errors are computed for the adjusted and original noisy time series, respectively. The quantity  $R$  is a measure of the self-consistency of the time series. (In other words,  $R$  measures how much better on the average the output attractor obeys Eq. 3 as one hops from point to point.)

When 1% noise is added to the input as described above, the noise reduction (measured with the actual map) is 79%.<sup>10</sup> Nearly identical results are obtained when the input contains only 0.1% noise. In addition, noise levels can be reduced almost as much in cases where the noise is added to the dynamics, i.e., where the input is of the form  $\{x_{i+1}: x_{i+1} = f(x_i + \eta_i, x_{i-1} + \eta_{i-1}), \eta_i, \eta_{i-1} \text{ random}\}$ . When the program is run on noiseless input, the mean error in the output is 0.025% of the attractor extent, which suggests that errors arising from small nonlinearities are negligible when the input contains enough points.

## 7 Simplicial Approximations of Dynamical Systems

Recent work has shown that simplicial approximations of dynamical systems can reproduce the behavior of the original system to high accuracy [34]. (See also [33] for a bilinear approach.) In particular, the fractal structure of the original attractors and basin boundaries is preserved over many scales. Such approximations can yield significant computational savings, especially when the original system consists of ordinary differential equations.

---

<sup>10</sup>The pointwise error is measured using Eq. 3. However, the attractor can be embedded in more than two dimensions when performing the noise reduction.

This approach can be extended in a natural way to generate simplicial approximations of the dynamics on attractors reconstructed from experimental data. Our objective here is to find an approximate dynamical system in a neighborhood of the attractor as follows.

A *simplex* in an  $m$  dimensional space is a triangle with  $m + 1$  vertices. Suppose the map is known at each point on a grid. Then there is a unique way to extend the map linearly to the interior of the simplex  $S$  whose vertices are grid points. Given a point  $P$  in the interior of  $S$ , let  $\{b_i\}_{i=0}^m$  be its corresponding *barycentric coordinates* (see [34] for an algorithm to compute them). Let  $f(v_i)$  be the map at the  $i$ th vertex. The dynamical system at  $P$  is iterated by computing

$$\Phi(P) = \sum_{i=0}^m b_i f(v_i). \quad (4)$$

We apply this method to experimental data by finding a linear approximation of the dynamics at each vertex  $v_i$  with the least squares method described above, using a collection of points in a small ball around  $v_i$ . The maps are stored and retrieved using a hashing algorithm similar to that described in [34]. This yields a piecewise linear approximation of the dynamics from a set of experimental data which can be analyzed with the methods that previously were available only to theorists.<sup>11</sup>

We illustrate the approach using a time series of 32,768 values from the Hénon map with  $\alpha = 1.2$ ,  $\beta = 0.3$  using Eq. 3 and adding 0.1% noise as described above. The original attractor is shown in Fig. 6(a). We take a grid of points which are spaced at 1% intervals (this and subsequent distances are expressed as a fraction of the original attractor extent). The time series is embedded in two dimensions, and a linear approximation of the dynamics is computed at each grid point for which 50 or more attractor points can be collected with a ball of radius 0.03; the set of such grid points is shown in Fig. 6(b). We take an initial condition near the original attractor and show the first 3000 iterates using Eq. 4 in Fig. 6(c). Although some defects are visible, the attractor produced by the approximate dynamical system looks almost identical to the original one.

---

<sup>11</sup>This approach is less ambitious than that of Crutchfield and McNamara [7], who attempt to find a single set of nonlinear difference equations that creates the observed attractor.

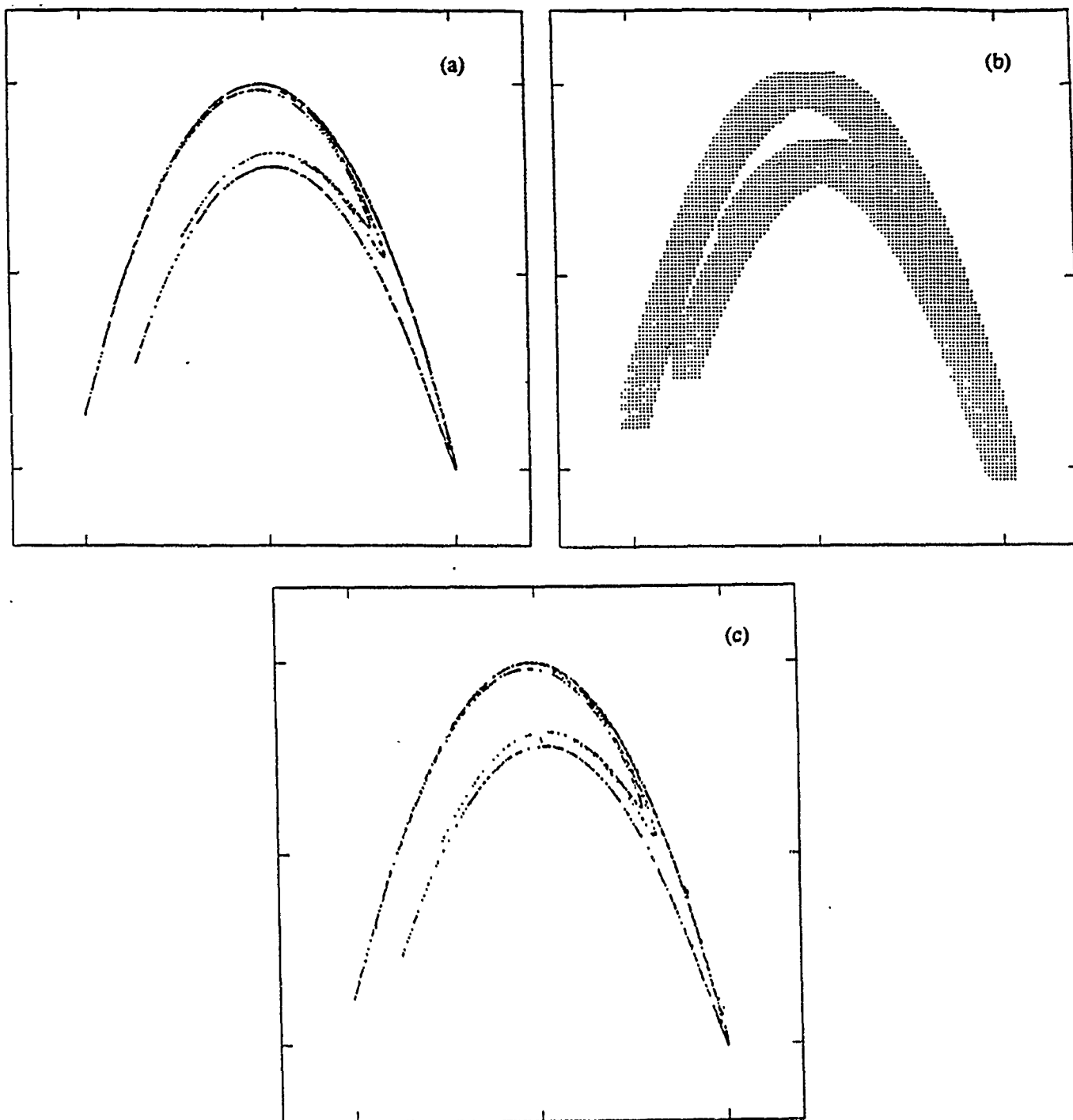


Figure 6: (a) Hénon attractor computed from Eq. 3 with  $\alpha = 1.2$ ,  $\beta = 0.3$ . (b) 1% grid on which linear approximations of the dynamics are computed from the available attractor points. (c) Attractor produced by the simplicial approximations.

period	$D = 2$	exact	$D = 3$
1	1.793	1.695	1.757
2	2.178	2.199	2.183
4	4.226	4.329	4.051
6	10.38	10.70	9.626
6	10.38	11.32	12.12
8	25.80	24.88	30.25
8	20.02	20.60	20.38
8	17.70	24.32	21.70

Table 1. The largest eigenvalues of the Jacobian of the periodic orbits located using the simplicial approximation of the Hénon attractor.

One application of simplicial approximations is the location of periodic saddles and the estimation of the largest eigenvalue of the corresponding Jacobian. That is, if  $x$  is a periodic point of period  $p$ , then we find the eigenvalue of  $Df^p(x)$  of largest modulus, where  $Df^p(x)$  refers to the matrix of partial derivatives of the  $p$ th iterate of the map  $f$  evaluated at  $x$ .

Given an initial guess for  $x$ , one can apply Newton's method using the maps computed at the grid points and Eq. 4 to locate the saddle using the simplicial approximations. Likewise, Eq. 3 can be used to locate the corresponding "exact" saddle. Saddle orbits up to period 8 have been computed in this way. In all cases, the saddle point for the simplicial approximation is within 2% of the corresponding saddle point for the Hénon map. Table 1 shows the largest eigenvalues of the saddle orbits. (The columns labeled  $m = 2$  and  $m = 3$  refer to the embedding dimension used to reconstruct the attractor.) In most cases, the relative error is only a few percent, and in no case exceeds 25%. (The largest relative error is for the period 8 saddles, where one finds the eigenvalue of the product of 8 Jacobians computed from the least squares.)

This method can be extended to experimental data sets. However, there are relatively stringent requirements on the data that can be handled: the time series must be long enough to trace out many trajectories near the principal unstable saddle orbits, and the noise level must be low. (Presumably noisy data can be preprocessed using the approach described in Section 3.)

The current computer implementation uses a large amount of disk space to store the linear map approximations at the grid points.

We have constructed a simplicial approximation for an attractor obtained from a Belousov-Zhabotinskii chemical reaction [6, 28]. The attractor is reconstructed in three dimensions from a set of 32,768 measurements of bromide ion concentration. The phase portrait is shown in 7(a).

Linear approximations of the dynamics are computed at each point of a grid consisting of 50 intervals along each coordinate axis for which 50 or more attractor points can be located within an 8% radius of the grid point. This produces a database of 59,550 maps. We observe from graphical evidence that many trajectories approach what appears to be a period 3 saddle in the middle of the attractor. Using initial guesses from some of the trajectories, we apply Newton's method to locate the saddle orbit shown in Fig. 7(b). Moreover, we obtain estimates of the Jacobian  $DF$  of the map evaluated at a point on saddle orbit. The eigenvalues of  $DF$  are estimated as  $\lambda_1 = 1.14$ ,  $\lambda_2 = 0.102$ , and  $\lambda_3 = -1.53$ . These quantitative results confirm that the orbit is a saddle since  $\lambda_1 > 0 > \lambda_3$ . (Note that one expects  $\lambda_2 = 0$  for a flow generated from a set of differential equations.)

## 8 Conclusion

Methods for approximating the dynamics of attractors reconstructed from experimental data provide powerful tools. Most of the same procedures that have been so important for theoretical insight, such as Poincaré maps, unstable fixed points and their manifolds, basin boundaries, and the like, are now available to experimenters, at least in cases where the dynamics are low dimensional. There is little doubt that these tools will lead to breakthroughs in the understanding of a wide variety of physical systems. However, considerable effort is needed before we learn which kinds of systems will benefit most from these types of analyses. Significant improvements in technique will certainly extend the applicability of dynamical embedding methods, for example to higher dimensional attractors.

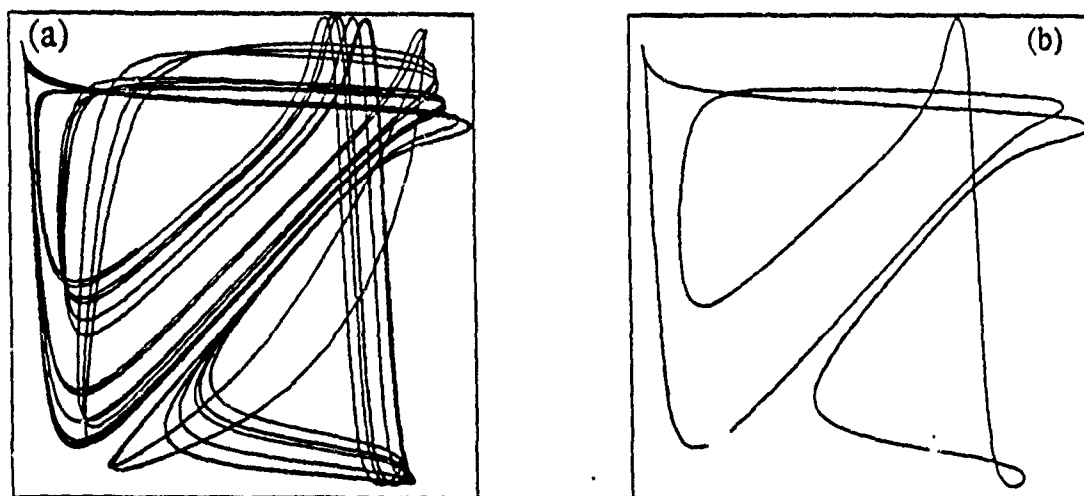


Figure 7: (a) The attractor reconstructed from a time series of bromide ion concentrations in a Belousov-Zhabotinskii chemical reaction. (b) The period 3 saddle orbit.

## Appendix

In this appendix we outline a possible alternative noise reduction method based on the theory of least squares when all the quantities in the regression are measured with error.

In ordinary least squares, the variables in the problem fall into two classes: the *independent* variables, which are known exactly, and the *dependent* variables, which are observations assumed to be functions of the independent variables. The dependent variables are subject to random errors that are assumed independent and identically distributed (i.i.d.).

On an attractor reconstructed from experimental data, we assume that the mapping which takes points in a sufficiently small ball to their images is approximately linear. However, the locations of all the points are subject to small random errors because of the noise. Hence one cannot describe the points as independent variables and their images as dependent variables. The usual least squares method produces a biased estimate of the linear map, and this bias does not decrease if more observations are added [15, 10].

The so-called "errors in variables" least squares methods can be used to handle the latter problem. This approach can be used to obtain both an estimate of the linear map as well as estimates of the "true" values of each of the observations.

At first this appears to be an underdetermined problem: from  $n$  pairs of observations one wants to compute the parameters of the functional relation between them as well as estimates of the  $n$  actual pairs.<sup>12</sup> However, it is possible to solve this problem by making some assumptions about the errors [15, 10].

In our case, we assume that the errors in the location of each point and its image are i.i.d. In particular, we let the covariance matrix of the errors in the variables be the identity matrix. This assumption is valid whenever the noise is independent of the dynamics.<sup>13</sup>

We illustrate the procedure for the case where we are given a collection of  $n$  points (in  $\mathbf{R}^m$ ) and their images. Following Jefferys [13], we form a set

---

<sup>12</sup>In the statistical literature, the problem is said to be *unidentified*.

<sup>13</sup>Dynamical noise (i.e., each point is perturbed slightly before iterating) yields a covariance matrix which depends on the point. However, as long as the dynamical noise is small, our assumptions about the covariance matrix of the errors should not compromise the accuracy of the method.

of  $n$  equations of condition given by

$$f_i(\mathbf{x}_i) = \mathbf{x}_{n+i} - A\mathbf{x}_i - \mathbf{b}_i \equiv \mathbf{x}_{n+i} - L(\mathbf{x}_i) \quad (5)$$

where  $\mathbf{x}_i$  is the  $i$ th point,  $\mathbf{x}_{n+i}$  is its observed image,  $A$  is an  $m \times m$  matrix, and  $\mathbf{b}$  is an  $m$ -vector. The goal is to find estimates of  $L$  (i.e.,  $A$  and  $\mathbf{b}$ ), together with perturbations  $\hat{\mathbf{v}}$ , such that

$$f_i(\mathbf{x}_i + \hat{\mathbf{v}}_i) = (\mathbf{x}_{n+i} + \hat{\mathbf{v}}_{n+i}) - L(\mathbf{x}_i + \hat{\mathbf{v}}_i) = 0$$

and such that the quadratic form

$$S_0 = \frac{1}{2} \hat{\mathbf{v}}^t \sigma^{-1} \hat{\mathbf{v}} \quad (6)$$

is minimized. The superscript  $t$  denotes transpose and  $\sigma$  is the covariance matrix of the observations (which we assume is the identity matrix here).

This minimization problem can be solved using Lagrange multipliers (see [13] and [14] for a numerical algorithm). The solution gives  $A$  and  $\mathbf{b}$  together with estimates  $\mathbf{x}_i + \hat{\mathbf{v}}_i$  of the "true" observations. It can be shown [10] under fairly mild hypotheses that the estimates of  $L$  and the observations are the best in the class of linear estimators.

One way to approach noise reduction is to extend Eq. 5 to include several iterations of the observed points. Given a collection of points in a ball, together with the next  $p$  iterates of each point, the method above is used to find a collection of linear maps  $L_1, L_2, \dots, L_p$  approximating the dynamics. The method also finds estimates of the actual observations. In this approach, therefore, the calculation of the maps and the adjustment of the trajectories is done in one step. Moreover, each point and its image exactly satisfy a linear relationship.

Of course,  $p$  cannot be too large, because nonlinear effects eventually will become significant when the dynamics are chaotic. On the other hand, Eq. 5 provides a natural way to include quadratic or other nonlinear terms.

We have written a computer program to implement this alternative noise reduction algorithm. So far, the results of this approach have not been as good as those from the method described in the main part of the paper, but further refinement should improve them.

## Acknowledgments

Dan Lathrop provided invaluable assistance in finding periodic orbits in the Hénon and BZ attractors. We thank Bill Jefferys for useful discussions and computer software for the errors in variables least squares problem. Andy Fraser, Randy Tagg and Harry Swinney all provided helpful suggestions. This research is supported by the Applied and Computational Mathematics Program of the Defense Advanced Research Projects Agency (DARPA-ACMP) and by the Department of Energy Office of Basic Energy Sciences.

## References

- [1] R. P. Behringer and G. Ahlers, *J. Fluid Mech.* **125** (1982), 219; G. Ahlers and R. P. Behringer, *Phys. Rev. Lett.* **40** (1978), 712.
- [2] J. L. Bentley and J. H. Friedman, *ACM Comput. Surv.* **11** (1979), 397.
- [3] See for example S. M. Bozic, *Digital and Kalman Filtering* (London: Edward Arnold Publishers Ltd., 1979).
- [4] A. Brandstätter and H. L. Swinney, *Phys. Rev. A* **35** (1987), 2207.
- [5] M. Casdagli, "Nonlinear Prediction of Chaotic Time Series," preprint (Dec. 1987).
- [6] K. C. Coffman, Ph.D. thesis, University of Texas at Austin, 1987.
- [7] J. P. Crutchfield and B. McNamara, *Complex Systems* **1** (1987), 417.
- [8] J. J. Dongarra, C. B. Moler, J. R. Bunch, and G. W. Stewart, *LINPACK User's Guide* (Philadelphia: Society for Industrial and Applied Mathematics, 1979).
- [9] J.-P. Eckmann and D. Ruelle, *Rev. Mod. Phys.* **57** (1985), 617.
- [10] W. A. Fuller, *Measurement Error Models* (New York: John Wiley & Sons, 1987).
- [11] S. M. Hammel, J. A. Yorke, and C. Grebogi, *J. of Complexity* **3** (1987), 136; *ibid.*, *Bull. Amer. Math. Soc.* **19** (1988), 465.

- [12] D. Hirst, Ph.D. dissertation, University of Texas, Dec. 1987.
- [13] W. H. Jefferys, *Astron. J.* **85** (1980), 177.
- [14] W. H. Jefferys, *Astron. J.* **86** (1981), 149.
- [15] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 2 (London: Charles Griffin & Company Limited, 1961), p. 375.
- [16] A. Libchaber, S. Fauve, and C. Laroche, *Physica D* **7** (1983), 73.
- [17] J.-P. Eckmann, S. O. Kamphorst, D. Ruelle and S. Ciliberto, *Phys. Rev. A* **34** (1986), 4971.
- [18] J. D. Farmer and J. J. Sidorowich, *Phys. Rev. Lett.* **59** (1987), 845.
- [19] P. R. Fenstermacher, H. L. Swinney, and J. P. Gollub, *J. Fluid Mech.* **94** (1979), 103.
- [20] W. Franceschini and L. Russo, *J. Stat. Phys.* **25** (1981), 757.
- [21] A. Fraser and H. L. Swinney, *Phys. Rev. A* **34** (1986), 1134.
- [22] E. G. Gwinn and R. M. Westervelt, *Phys. Rev. A* **33** (1986), 4143.
- [23] M. Hénon, *Comm. Math. Phys.* **50** (1976), 69.
- [24] E. N. Lorenz, *J. Atmos. Sci.* **20** (1963), 130.
- [25] S. W. MacDonald, C. Grebogi, E. Ott and J. A. Yorke, *Physica D* **17** (1985), 125.
- [26] For example, see the papers in *Dimensions and Entropies in Chaotic Systems*, ed. by G. Mayer-Kress (Berlin: Springer-Verlag, 1986), and references therein.
- [27] For example, see L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing* (Englewood Cliffs, N. J.: Prentice-Hall, 1975).
- [28] J.-C. Roux, *Physica D* **7** (1983), 57; J.-C. Roux, R. H. Simoyi, and H. L. Swinney, *Physica D* **8** (1983), 257.

- [29] M. Sano and Y. Sawada, *Phys. Rev. Lett.* **55** (1985), 1082.
- [30] C. Sparrow, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors* (New York: Springer-Verlag, 1982).
- [31] R. Tagg, private communication.
- [32] F. Takens, in *Dynamical Systems and Turbulence*, ed. by D. A. Rand and L.-S. Young, *Springer Lecture Notes in Mathematics*, Vol. 898 (New York: Springer-Verlag, 1980), p. 366.
- [33] B. H. Tongue, *Physica D* **28** (1987), 401.
- [34] F. Varosi, C. Grebogi, and J. A. Yorke, *Phys. Lett. A* **124** (1987), 59.
- [35] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano, *Physica D* **16** (1985), 285.

November 1989

ACCESSIBLE SADDLES ON  
FRACTAL BASIN BOUNDARIES

by

Kathleen T. Alligood<sup>1</sup>  
Department of Mathematics  
George Mason University  
Fairfax, VA 22030

and

James A. Yorke<sup>1,2</sup>  
Department of Mathematics and  
Institute for Physical Science and Technology  
University of Maryland  
College Park, MD 20742

<sup>1</sup>Research partially funded by a contract from the Applied and Computational Mathematics Program, DARPA.

<sup>2</sup>Research partially funded by a grant from AFOSR.

## ABSTRACT

For a homeomorphism of the plane, the basin of attraction of a fixed point attractor is open, connected, and simply-connected, and hence is homeomorphic to an open disk. The basin boundary, however, need not be homeomorphic to a circle. When it is not, it can contain periodic orbits of infinitely many different periods.

Certain points on the basin boundary are distinguished by being accessible (by a path) from the interior of the basin. For an orientation-preserving homeomorphism, the accessible boundary points have a well-defined rotation number. We prove that this rotation number is rational if and only if there are accessible periodic orbits. In particular, if the rotation number is the reduced fraction  $p/q$ , then every accessible periodic orbit has minimum period  $q$ . In addition, if the periodic orbits are hyperbolic, then every accessible point is on the stable manifold of an accessible periodic point.

## 1. Introduction and Statement of Main Theorems

When a dynamical system has more than one attractor, the boundaries between respective basins of attraction can exhibit very complicated patterns. For invertible maps of the plane, these boundaries can be smooth or fractal, and they can contain infinitely many saddle-type periodic orbits. (By fractal, we mean that the set has non-integer Hausdorff dimension.) Two basins of attraction of the time  $2\pi$  map of the forced damped pendulum equation are shown in black and white in Figure 1. This picture was constructed by choosing a  $960 \times 520$  grid and, using each grid point as an initial condition, testing where its trajectory goes. The system has two fixed point attractors--one in the white region to which all grid points colored white tend under iteration by the map, and one in the black region to which all grid points colored black tend. The boundary between the black and white basins is fractal, making final state predictability very difficult. In addition, buried within the fractal layers of the boundary are saddle periodic orbits of arbitrarily high periods.

Even though the dynamics on the boundary appear to be very complicated, it has been observed (see, for example, [GOY]) that some points on the boundary exhibit regular behavior. We say that a point  $p$  on the boundary of an open set  $W$  is accessible from  $W$  if there is a path beginning in  $W$  such that  $p$  is the first boundary point which the path hits. Surprisingly, when the boundary is fractal, most points are not accessible. For the map in Fig. 1, there are two points that are saddles of period two (i.e., one period two orbit)

which are accessible from the white region, and all other points which are accessible from the white region are on the stable manifold of this periodic orbit. In this paper, we investigate the dynamics of the accessible points on basin boundaries. The paper is strongly motivated by numerical studies that repeatedly conclude there are accessible periodic saddles in the boundary. In fact, we know of no natural case of an area-contracting diffeomorphism having a basin boundary without accessible periodic orbits.

We would like to thank J. Mather and H. Nusse for helpful discussions.

Throughout this paper,  $W$  is a connected, simply-connected open set either in the plane  $\mathbb{R}^2$  or in the sphere  $S^2$ , and  $F$  is a homeomorphism (or diffeomorphism, if differentiability is required) of the plane or the sphere. We assume that  $W$  is invariant under  $F$ , (i.e.,  $F(W) = W$ ). Our main examples of such sets will be basins of attraction. In particular, the basin of attraction of an attracting fixed point must be such a set. (See Sec. 2.) We assume in addition that  $W$  is not the entire plane, in which case its boundary  $\partial W$  is more than one point, ~~and is a connected set~~. Since  $W$  is invariant under  $F$ ,  $\partial W$  is also an invariant set. All connected, simply-connected open sets are homeomorphic to an open disk. On the other hand, the boundary of such a region does not have to be topologically a circle, and examples abound in which the boundary of a basin of attraction is a fractal set. The characterization of a set  $W$

as a topological open disk occurs in the study of the Riemann Mapping Theorem which says that for any such set  $W$  there is always a one-to-one analytic map  $h$  of an open disk  $D$  onto  $W$ . The knowledge that the basin is topologically an open disk tells us nothing about the boundary of a basin, and it is our objective to describe the dynamics on the points in  $\partial W$  that are accessible from  $W$ .

In the following we say that  $p$  is accessible only if it is a point of  $\partial W$  that is accessible from  $W$ .

Caratheodory [C] investigated the behavior of the map  $h$  in the Riemann mapping theorem to see when  $h$  could be defined at boundary points of the disk. If  $\Gamma$  is a (continuous) path in  $W$  which limits on the accessible point  $p$ , then  $h^{-1}(\Gamma)$  is a (continuous) path in  $D$  limiting on exactly one point  $r$  in  $S^1$ , the boundary of  $D$ . We call such points as  $r$  trivial circle points; we call all other points on the circle non-trivial circle points. Caratheodory's approach was to construct a compactification of  $W$  which is topologically identical to  $\bar{D}$ , the closed disk. (His is not the standard compactification; points in this compactification which correspond to points in the boundary  $S^1$  of  $\bar{D}$  are called "prime ends" and are defined precisely in Sec. 5.)

We define a map  $h_c$  on points in  $D$  and on those points in  $\bar{D}$  that are trivial circle points by  $h_c(x) = h(x)$  for  $x$  in  $D$ , and  $h_c(r) = p$  where  $p$  is an accessible point and  $r$  is an associated trivial circle point, as defined above. It is clear from the construction that each accessible point is the image of at least one trivial circle point.

The map  $h_c$  is not necessarily one-to-one on trivial circle points.

(See Sec. 7; in particular, see Fig. 7.) However, once a path  $\Gamma$  in  $W$  limiting on an accessible point  $p$  is specified, then there is exactly one trivial circle point  $x$  which is the limit of  $h^{-1}(\Gamma)$ .

We mention two properties of accessible points and the map  $h_c$ :

PROPERTY 1 (DENSITY) The set of accessible points is dense in  $\partial W$ ; the set of trivial circle points is dense in  $S^1$ , the boundary of  $D$ .

PROPERTY 2 (EXISTENCE OF AN INDUCED MAP) There is a map, denoted  $f$  and called the induced map, from  $\bar{D}$  to itself such that  $h_c(f(x)) = F(h_c(x))$  when  $x$  is in  $D$  or  $x$  is a trivial circle point.

If  $p$  is an accessible point and  $\Gamma$  is a path in  $W$  ending at  $p$ , then  $F(\Gamma)$  is a path in  $W$  ending at  $F(p)$ . Hence, accessible points map to accessible points. It follows that  $f$  maps trivial circle points to trivial circle points. On the set of trivial circle points,  $f$  is one-to-one, onto, and order-preserving. Such a map can be uniquely extended to a homeomorphism defined on all of  $S^1$ .

These properties allow us to study the dynamical system on the closed disk, maintaining the dynamics on the accessible points. Since in general  $\partial W$  will include much more than the accessible points, much of  $\partial W$  is lost in this representation. For us, however, the simplification is advantageous since we wish to describe the dynamics

on the accessible points.

We have important examples in which  $W$  is not a basin even though a dense set of points in  $W$  have trajectories tending to an attractor. The following definition allows the inclusion of such examples. We say that  $\partial W$  is unstable in  $W$  if there is a neighborhood  $B_\epsilon$  of  $\partial W$  with the property that the set of points in  $B_\epsilon$  whose orbits eventually leave  $B_\epsilon$  is dense in  $B_\epsilon \cap W$ . (I.e., there is a dense set  $Q$  in  $B_\epsilon \cap W$  such that  $x \in Q$  implies that  $F^n(x)$  is in  $W \setminus B_\epsilon$  for some  $n > 0$ .) This definition is easily seen to be satisfied when  $W$  is a basin of attraction. It is also satisfied in the very different case where there is a dense orbit in  $W$ .

Certain types of periodic orbits in  $S^1$  merit particular attention. Let  $p \in S^1$  be a periodic point of period  $k$ . We say  $p$  is attracting on at least one side (of  $S^1$ ) if there exists  $x \in S^1$  such that  $x \neq p$  and  $\lim_{n \rightarrow \infty} f^{nk}(x) = p$ .

The following key theorem is proved in Sec. 5:

**THEOREM 1.1 (ATTRACTING LEMMA).** Assume that  $\partial W$  unstable in  $W$  and that for each  $k$  the fixed points of  $F^k$  are isolated. Then each periodic circle point that is attracting on at least one side is a trivial circle point.

An orientation preserving homeomorphism of the circle can be classified according to its rotation number--a number  $\rho$ , with  $0 \leq \rho < 1$ , which represents the average rotation of points under the

map. (A precise definition is given in Sec. 5.) The rotation number is independent of the choice of point on  $S^1$ . The idea of associating a single rotation number with each orientation preserving homeomorphism of the circle originated with Poincaré. Such a homeomorphism will have a periodic point if and only if its rotation number is rational. It will have a fixed point if and only if its rotation number is 0. We define the rotation number  $\rho(\partial W, F)$  of  $F$  on the accessible points of  $\partial W$  to be the rotation number of the induced map  $f$  on  $S^1$ . If  $W$  is a connected, simply-connected open set in  $\mathbb{R}^2$ , if  $F$  is orientation preserving, and if the closure of  $W$  is invariant under  $F$ , then  $W$  has a rotation number. In particular, if  $p$  is an isolated, attracting fixed point in  $\mathbb{R}^2$ , if its basin  $W$  is not all of  $\mathbb{R}^2$ , and if  $F$  is orientation preserving, then  $\partial W$  has a rotation number. (See Sec. 2.)

G.D. Birkhoff recognized that the set of accessible points is dense in the boundary of an invariant region and that their dynamics can be characterized by their rotation number. He used this idea in [B] to construct a map of the annulus into itself with an unusual invariant set  $J$ . On one hand,  $J$  resembles a closed Jordan curve in that each of its points is on the boundary of both an interior region  $S_{\text{int}}$  (containing one boundary circle of the annulus) and an exterior region  $S_{\text{ext}}$  (containing the other boundary circle). On the other hand,  $J$  is "remarkable" in the sense that it contains a dense set of points accessible from  $S_{\text{int}}$  with one rotation number and a dense set accessible from  $S_{\text{ext}}$  with a different rotation number. To compare this situation with our hypotheses, notice that such a map has an

inverse for which  $J$  is unstable (in  $S_{\text{int}}$  and in  $S_{\text{ext}}$ ) and  $J$  is the boundary between the points which go outward and those which go inward (under the inverse).

Cartwright and Littlewood further developed these ideas in [C-L1], where they prove the existence of and determine the stability of periodic orbits for a certain class of second order differential equations in the plane. More recently, J. Mather has given purely topological proofs of some of the topological results of Carathéodory in [M1] and has used the theory to study invariant sets for area-preserving homeomorphisms of the annulus [M2], [M3]. We rely on the proofs in the above references of Cartwright-Littlewood and Mather for much of the material on prime ends given in Secs. 5, 6, and 7. A general reference for Carathéodory's theory is [C-Lo], Chapter 9.

The following argument explains the significance of the Attracting Lemma. Assume that the rotation number of  $f$  on  $S^1$  is rational (say the reduced fraction  $p/q$ ). Then  $S^1$  will have at least 1 fixed point under  $f^q$ , (i.e., a periodic point of period  $q$ ). If a trivial circle point  $x$  is not fixed under  $f^q$ , then its orbit converges to a fixed point  $r$  under iterates of  $f^q$ . By the Attracting Lemma,  $r$  is necessarily a trivial circle point. Corresponding to  $r$  is an accessible point  $p$  on  $\partial W$ . By Property 2,  $p$  is fixed under  $F^q$ . Thus we have the following result:

THEOREM 1.2. Assume that  $\partial W$  is unstable in  $W$  and that for each  $k$  the fixed points of  $F^k$  are isolated. Assume further that the rotation number  $\rho(\partial W, F)$  is  $p/q$  (resp., 0). Then there is an accessible fixed point of  $F^q$  (resp.,  $F$ ) on  $\partial W$ .

In Sections 6 and 7 we describe the dynamics on the set of accessible points under the hypotheses that  $\rho$  is rational,  $F$  is a diffeomorphism, and periodic points in the boundary are hyperbolic. (A periodic point  $p$  is hyperbolic if the Jacobian matrix  $DF(p)$  has no eigenvalues with absolute value 1.) By the Inverse Function Theorem, a hyperbolic point is isolated from other periodic points of the same period (or smaller period). In the following theorem, which is a special case of Theorem 6.1 in Sec. 6, we assume that  $W$  is a basin of attraction: i.e., there exists a compact set  $K$  in  $W$  such that the " $\omega$ -limit set" of the orbit of each point  $x$  in  $W$  is non-empty and is contained in  $K$ . (Given a point  $x$ , the point  $z$  is in the  $\omega$ -limit set of the orbit of  $x$ , if there exists a sequence  $\{t_n\}$ , with  $t_n \rightarrow \infty$ , such that  $f^{t_n}(x) \rightarrow z$ .) If the orbit of each point in  $W$  is bounded, then there exists a compact set  $K' \subseteq K$  which is Liapunov stable (see Sec. 2 for definition) [BS].

THEOREM 6.1'. Assume that the periodic points of  $F$  in  $\partial W$  are hyperbolic and that  $W$  is a basin of attraction. If the rotation number  $\rho$  is rational, then every accessible point either is a periodic point or is in the stable manifold of an accessible periodic point.

Theorems 1.2 and 6.1' do not mention the *minimum* period of an accessible periodic orbit. Degeneracies can occur due to the fact that the map  $h_c$  is not necessarily one-to-one on trivial circle points, so that the period of the accessible points can strictly divide the period of the orbit on  $S^1$ . In Sec. 7 we prove that such degeneracies are ruled out for homeomorphisms of the plane, although they can still occur for homeomorphisms of the sphere. We use the following two results. The first, a converse of Theorem 1.2 for planar maps, implies that the period of an accessible periodic point cannot be strictly smaller than the period of a trivial periodic circle point. The second implies that it cannot be strictly larger.

PROPOSITION 7.3. Let  $F$  be a homeomorphism of the plane  $\mathbb{R}^2$ . If there exists an accessible fixed point on  $\partial W$ , then  $\rho(\partial W, F)$  is 0.

PROPOSITION 7.4. If  $\rho = 0$ , then every accessible periodic point in  $\partial W$  is a fixed point.

COROLLARY 7.5. Let  $F$  be a homeomorphism of the plane  $\mathbb{R}^2$ . If  $\rho \neq 0$  is the reduced fraction  $p/q$ , then every accessible periodic point in  $\partial W$  has minimum period  $q$ .

The next corollary (a special case of Cor. 7.6) follows, although not directly, from Prop. 7.3, Prop. 7.4, and Thm. 6.1'. In particular, it remains to be shown that if the orbit of an accessible point converges to a fixed point in  $\partial W$ , then the fixed point is accessible. We point out that this corollary does not mention the rotation number  $\rho$ .

COROLLARY 7.6'. Assume the following conditions hold:

- (1)  $F$  is a diffeomorphism of the plane  $\mathbb{R}^2$ ;
- (2) the periodic points of  $F$  in  $\partial W$  are hyperbolic;
- (3)  $W$  is a basin of attraction; and
- (4) either (i) there exists an accessible period point of minimum period  $q$ , or  
(ii) there exists an accessible point which converges (under  $f^q$ ) to a periodic point of minimum period  $q$ .

Then every accessible point in  $\partial W$  either is a periodic point of minimum period  $q$  or is in the stable manifold of such a periodic point.

In Sec. 2 we define a general class of connected, compact attractors and show that attractors in this class have connected, simply-connected basins. In Sec. 3 we study the orientation-reversing case, and in Sec. 4 we apply Theorem 1.2 to a class of chaotic attractors, viewed as boundaries for the inverse of the map  $F$ .

Figures 1 through 4 were made using Dynamics [Y].

## 2. Attractors with Simply-Connected Basins

If  $A$  is a hyperbolic fixed point, then  $A$  has a connected, simply-connected neighborhood which contracts to it under iteration by  $F$ . In this case, the entire basin of  $A$  (see Sec. 1 for definition) is connected and simply connected. Here we look at a more general class of attractors and show that their basins are connected and simply connected and thus satisfy the hypotheses of Theorems 1.1 and 1.2. (The hypothesis that the boundary  $\partial W$  is unstable in  $U$  is trivially satisfied if either the attractor  $A$  does not intersect  $\partial W$  or if  $A$  has a dense orbit and is not a subset of  $\partial W$ .)

For a closed set  $S$ , let  $S_\epsilon$  be the  $\epsilon$ -neighborhood of  $S$ ; i.e.,  $S_\epsilon$  is the set of points  $y$  such that  $\min_{x \in S} \|x - y\| < \epsilon$ , where  $\|\cdot\|$  denotes the Euclidean norm in  $\mathbb{R}^2$ . We say a set  $A$  is a regular attractor if  $A$  satisfies the following three properties:

(2.1)  $A$  is compact and connected;

(2.2)  $A$  is Liapunov stable; i.e., for each neighborhood  $Y$  of  $A$  there exists  $\epsilon > 0$  such that  $A_\epsilon \subset Y$ , and if  $x \in A_\epsilon$  then  $F^n(x) \in Y$ , for all  $n \geq 1$ ;

(2.3) The basin of  $A$  contains an open neighborhood of  $A$ .

In the following proposition, "area-contracting" means specifically that there exists a number  $\xi$ , where  $\xi < 1$ , such that  $|\det DF(x)| < \xi$ , for all  $x$  in  $\mathbb{R}^2$ .

PROPOSITION 2.4. Let  $F$  be an area-contracting map of the plane. If  $A$  is a regular attractor, then the basin  $U$  of  $A$  is open, connected, and simply connected.

*Proof.* Let  $Y \subset U$  be an open neighborhood of  $A$ . Let  $\varepsilon > 0$  be given such that  $F^n(A_\varepsilon) \subset Y$  for all  $n \geq 0$ . Select  $\delta$ ,  $0 < \delta < \varepsilon$ , such that  $F^n(A_\delta) \subset A_\varepsilon \subset Y \subset U$ , for all  $n \geq 1$ . Such  $\varepsilon$  and  $\delta$  exist, since  $A$  is Liapunov stable.

Let  $x \in U$  be given. Choose  $k > 0$  such that  $F^k(x) \in A_\varepsilon$ . Since  $A_\varepsilon$  is open, there is an open neighborhood  $V_x$  of  $x$  which maps into  $A_\varepsilon$  under  $F^k$ . Thus each point  $x$  in  $U$  has an open neighborhood  $V_x$  in  $U$ , and  $U$  is open.

Let  $x_1$  and  $x_2$  in  $U$  be given. Choose integers  $P > 0$  and  $Q > 0$  such that  $F^P(x_1) \in A_\delta$  and  $F^Q(x_2) \in A_\delta$ . Define  $m = \max \{P, Q\}$ . Then  $F^m(x_1)$  and  $F^m(x_2)$  are in  $A_\varepsilon$ . Since  $A$  is connected,  $A_\varepsilon$  is connected for each  $\varepsilon > 0$ . Hence,  $A_\varepsilon$  is an open, connected set. Since open, connected sets are path connected, there is a path  $\Gamma$  in  $Y$  connecting  $F^m(x_1)$  and  $F^m(x_2)$ . Thus  $F^{-m}(\Gamma)$  lies in  $U$  and connects  $x_1$  to  $x_2$ . Therefore,  $U$  is connected.

It remains to show that  $U$  is simply connected. Suppose that  $U$  is not simply connected, and let  $C$  be a simple closed curve in  $U$  which bounds a region  $D$  containing a set  $S$  (consisting of one or more points) that is not in  $U$ . This implies that the distance between  $F^n(S)$  and  $A$  is at least  $\varepsilon$ , for all  $n \geq 0$ . Select  $\alpha$ ,  $0 < \alpha < \delta$ , such that  $F^n(A_\alpha) \subset A_\delta$ , for all  $n \geq 0$ . Since  $C$  is compact, there exists

an integer  $j(\alpha) > 0$  such that  $F^{j(\alpha)}(C) \subset A_\alpha$ . Therefore,  
 $F^j(C) \subset A_\delta$ , for all  $j \geq j(\alpha)$ . We conclude that the distance between  
 $F^n(S)$  and  $F^n(C)$  is at least  $\varepsilon - \delta$  for all  $n \geq j(\alpha)$ . On the other hand,  
 since  $F$  is area contracting, the distance between  $F^n(S)$  and  $F^n(C)$   
 converges to zero as  $n \rightarrow \infty$ . This contradicts the fact that  $\varepsilon - \delta > 0$ .  
 Therefore,  $U$  is simply connected. ■

### 3. Continuation and Orientation-Reversing Maps

Let  $F_\lambda$  be a homeomorphism of  $\mathbb{R}^2$  depending on a scalar parameter  $\lambda$ . We assume that  $F_\lambda$  has a fixed point regular attractor  $A_\lambda$ , which depends continuously on  $\lambda$ , for each  $\lambda$ . We define the maximal basin  $W_\lambda$  to be the largest open set having a dense set of points that are attracted to  $A_\lambda$  under  $F_\lambda$ . Let  $B_\lambda$  be the boundary of  $W_\lambda$ ; and let  $\rho_\lambda$  be the rotation number of  $F_\lambda$  on  $C_\lambda$ , the accessible points in  $B_\lambda$ . For a parametrized homeomorphism on a circle, the rotation number varies continuously with the parameter (see, for example, [D]). Unlike the circle case, however,  $\rho_\lambda$  is not necessarily continuous in  $\lambda$ . In fact, the boundary  $B_\lambda$  can jump discontinuously, even when there is no change in the attractor. It was shown in [HJ] (see also [GOY] and [ATY]) that when the stable and unstable manifolds of an accessible saddle on the boundary become tangent at  $\lambda = \lambda_*$  and then cross for  $\lambda > \lambda_*$ , the stable manifold jumps a positive distance  $\epsilon$  (not dependent on  $\lambda$ ) into  $W_{\lambda_*}$  for each  $\lambda > \lambda_*$ . Figure 2 shows in black the basin of attraction of infinity for three different values of the parameter  $\lambda$  in the Henon map

$$F_{\lambda,b}(x,y) = (\lambda - x^2 - by, x) \quad (3.1)$$

where  $b$  is fixed at 0.3. There is a period two attractor in the white region to which the orbits of almost all white points tend. Numerical experiments indicate that for  $\lambda=1.39$  (in Fig. 2a), a period-four saddle orbit and its stable manifold are the only boundary points

accessible from the white region. There is a tangency of the stable and unstable manifolds of consecutive points in this orbit at  $\lambda = \lambda_*$ ,  $\approx 1.395$ . Specifically, if we number the four points in the orbit  $x_1, \dots, x_4$  consecutively (in the counter-clockwise direction around the basin boundary), and if we set  $x_n = x_{n(\text{mod}4)}$  for  $n > 4$ , then  $\{x_n\}$  is a periodic trajectory. At  $\lambda = \lambda_*$  the unstable manifold of  $x_i$  is tangent to the stable manifold of  $x_{i+1}$ . For each  $\lambda > \lambda_*$ , black points appear in what was the interior of the white region. In addition, it has been numerically observed that for each  $\lambda > \lambda_*$  (near  $\lambda_*$ ), the set  $C_\lambda$  of accessible boundary points is composed of a period-three saddle and its stable manifold. Fig. 2bc show in black the basin of attraction of infinity at  $\lambda = 1.4$  and  $\lambda = 1.42$ , respectively, with the accessible period-three saddle. A numerical investigation of rotation numbers for the orientation-preserving, area-contracting Henon map appears in [AS].

When  $f$  is an orientation-reversing homeomorphism, the possible dynamics on accessible orbits are limited. For a connected, simply connected basin of attraction  $W$ , an orientation-reversing homeomorphism on  $\bar{W}$  restricts to an orientation reversing-homeomorphism on  $\partial W$ . Again, we study the dynamics on  $\partial W$  through its association with the circle. An orientation-reversing homeomorphism  $f$  of  $S^1$  must have fixed points. It may or may not have periodic points of period two. Notice, however, that  $f$  can have no periodic points of minimum period greater than 2. The map  $f^2$  is orientation preserving and has rotation number 0 since it has fixed points. But an orientation-preserving homeomorphism of the circle with rotation

number 0 has no periodic orbits of minimum period greater than 1.

Suppose  $f$  has a periodic orbit of minimum period  $k$ ,  $k \geq 3$ . Then  $f^2$  has a periodic orbit of minimum period  $k/2$ , if  $k$  is even, or of minimum period  $k$ , if  $k$  is odd. Thus  $f$  has only periodic points of period one or two.

We have the following restatements of Theorems 1.1 and 1.2 for orientation-reversing maps:

**THEOREM 3.2 (ATTRACTING LEMMA).** Let  $F$  be an orientation-reversing homeomorphism of the plane. Assume that  $\partial W$  is unstable in  $W$ . Assume further that the fixed points of  $F^2$  in  $\partial W$  are isolated. Then each circle point that is fixed under  $f^2$  and is attracting on at least one side is a trivial circle point.

**THEOREM 3.3.** Under the hypothesis of Theorem 3.2, there is an accessible fixed point of  $F^2$  on  $\partial W$ .

Let  $F_\lambda$  be a one-parameter family of orientation-reversing homeomorphisms. From Theorem 3.3, we observe that if a metamorphosis occurs for  $F_\lambda$ , then  $B_\lambda$  must jump to different fixed points of  $F^2$ .

Example. The Henon map (3.1) is orientation reversing for  $b < 0$ . It is easily verified that  $F_{\lambda, b}$  can have at most 2 fixed points and at most one periodic orbit of minimum period two. In this situation the possible metamorphoses are severely limited by Theorem 3.3. As long as the period-two orbit and one of the fixed points is in the attractor (and the hypotheses of Thm. 3.2 are satisfied), no metamorphoses will occur. If, however, the basin becomes disconnected, as shown in Fig. 3, then the theorem no longer applies and the boundary can be fractal. Fig. 3 shows in white the basin of a two-piece attractor (which is also plotted in the white region). A metamorphosis has occurred, and there is no longer an accessible fixed point on the boundary. Now the accessible saddle has period six.

The existence of periodic orbits in the maximal basin of the attractor but not in the attractor itself is also restricted by Thm. 3.2. Suppose (3.1) has a regular attractor  $A$  (i.e.,  $A$  satisfies Properties (2.1)-(2.3)). If  $A$  contains a fixed point and an orbit of period two or if (3.1) is in a parameter range where there is no period-two orbit and  $A$  contains a fixed point, then the basin  $U$  of  $A$  is necessarily bounded by the stable manifold of the (other) saddle fixed point  $p$ . (For every choice of parameter values, the orbits of some points in the plane go to infinity; thus the basin  $U$  has a boundary.) In particular, under these hypotheses, there are no *periodic orbits* in the region containing  $A$  and bounded by  $W^s(p)$  except those in  $A$ .

#### 4. Rotation Numbers for Chaotic Attractors.

Here we look at a class  $\mathcal{A}$  of non-periodic attractors in the plane: an attractor  $\Theta$  is in  $\mathcal{A}$  if  $\Theta$  is compact, connected, invariant under  $F$ , and contains more than one point. In order to apply Theorem 1.2, we show how to assign a rotation number to an attractor in the class  $\mathcal{A}$ , assuming that  $F$  is an area-contracting homeomorphism of the plane. This approach is reminiscent of Birkhoff [B] and also of Cartwright and Littlewood [C-L2] and Levinson [L] who studied attractors in forced van der Pol type equations.

In looking at the Poincaré map of such equations, Cartwright and Littlewood showed that there are invariant annuli which have unequal rotation numbers on the boundary circles and which possess strange attracting sets. Each such attractor is the boundary of the inside contracting and outside contracting parts of the annulus. The existence of different rotation numbers inherited from the boundary circles was evidence to them of a continuum attractor which was not homeomorphic to  $S^1$ . Levinson gave a careful analysis of the attracting invariant set of a piecewise-linear version of this map in [Ln]. His work set the stage for the discovery of the horseshoe map by Smale. See also Levi's analysis of forced van der Pol type equations in [Li].

Let  $Z = \mathbb{R}^2 \cup \{\infty\}$  be the one-point compactification of  $\mathbb{R}^2$ . Then  $F$  extends to a homeomorphism of  $Z$  by setting  $F(\{\infty\}) = \{\infty\}$ .

LEMMA 4.1. Let  $F$  be an area-contracting homeomorphism of the plane  $\mathbb{R}^2$ . If  $\Theta$  is in  $\mathcal{A}$ , then  $Z - \Theta$  is connected and simply-connected in  $Z$ .

*Proof.* Since  $\Theta$  is connected, each component of  $Z - \Theta$  is simply connected in  $Z$ . (This simple fact follows most clearly from Alexander Duality with Čech cohomology. See, for example, [Do].) Since  $\Theta$  is compact, only one component  $D_\infty$  of  $\mathbb{R}^2 - \Theta$  has infinite area (in  $\mathbb{R}^2$ ) and, given any bound  $\eta$ , there are only finitely many other components with area larger than  $\eta$ . Let  $D_M$  be a component of  $\mathbb{R}^2 - \Theta$  with maximum finite area in  $\mathbb{R}^2$ . Since  $F^{-1}$  is area-expanding and components of  $\mathbb{R}^2 - \Theta$  map onto other components of  $\mathbb{R}^2 - \Theta$ ,  $F^{-1}$  maps  $D_M$  onto  $D_\infty$ . But  $F^{-1}$  also maps  $D_\infty$  onto  $D_\infty$ , contradicting the fact that  $F^{-1}$  is a homeomorphism. Thus  $Z - \Theta$  is connected and simply connected in  $Z$ . ■

Now we can apply Theorem 1.2 to  $\Theta$ , which is the boundary of the open, connected, simply-connected region  $Z - \Theta$ . By looking at  $F^{-1}$  instead of  $F$ , it can be shown that  $\Theta$  is unstable in  $Z - \Theta$ , as follows. Let  $\Theta_\varepsilon$  be an  $\varepsilon$ -neighborhood of  $\Theta$ , and let  $D$  be an open set in  $\Theta_\varepsilon \cap (Z - \Theta)$ . Since  $F^{-1}$  is area-expanding, the area enclosed by the boundary of  $D$  becomes unbounded under iteration by  $F^{-1}$ . It can easily be shown that almost all points in  $D \cap \Theta_\varepsilon$  eventually will be mapped out of  $\Theta_\varepsilon$  under iteration of  $F^{-1}$ ; hence,  $\Theta$  is unstable in  $Z - \Theta$  under  $F^{-1}$ . Theorem 1.2 provides the following result:

PROPOSITION 4.2. Let  $F$  be an area-contracting homeomorphism of the plane, and let  $\Theta$  be in the class  $\mathcal{A}$  of attractors. Assume that, for each  $k$ , the fixed points of  $F^k$  are isolated. If the rotation number  $\rho(\Theta, F)$  is the reduced fraction  $p/q$ , then there is an accessible fixed point of  $F^q$  on  $\Theta$ .

Figure 4 shows an attractor for the Ikeda map with an accessible period 6 orbit. For a typical area-contracting diffeomorphism depending on a parameter  $\lambda$ , we conjecture that the rotation number  $\rho(\lambda)$  will vary continuously, except possibly at a discrete set of values of  $\lambda$ , and that  $\rho(\lambda)$  will be irrational for a non-empty set of  $\lambda$  of measure 0.

## 5. Proof of the Attracting Lemma.

Let  $F$  be an orientation-preserving homeomorphism of  $Z = \mathbb{R}^2 \cup \{\infty\}$ , the 1-point compactification of the plane. A simple arc  $Q$  in  $W$  with end points  $q_1$  and  $q_2$ ,  $q_1 \neq q_2$ , on  $\partial W$  and no other points on  $\partial W$  is called a crosscut of  $W$ . Each crosscut divides  $W$  into 2 subdomains, since  $W$  is simply connected. Let  $\{Q_n\}$  be a sequence of pairwise disjoint crosscuts such that  $Q_n$  separates  $Q_{n+1}$  from  $Q_{n-1}$ . Then there is a corresponding sequence  $\{V_n\}$  of subdomains of  $W$  such that  $V_n$  contains  $Q_{n+1}$  except for its endpoints. See Figure 5. The sequence  $V_1 \supset V_2 \supset V_3 \supset \dots$  is called a chain. If  $V = \{V_n\}$  and  $V' = \{V'_n\}$  are two chains, we say  $V$  divides  $V'$  if for each  $i$ , there is a  $j$  such that  $V'_j \subseteq V_i$ . We say  $V$  and  $V'$  are equivalent if each divides the other. Under this relation, an equivalence class of chains is called an end. A chain  $V$  is called prime if any chain which divides it is equivalent to it. A prime end is the equivalence class of a prime chain. For the unit disk  $D$  in  $\mathbb{R}^2$  a chain  $\{V_n\}$  is prime if and only if  $\bigcap_{n \in \mathbb{N}} \overline{V_n}$  is a single point (necessarily on the boundary  $S^1$ ). In general, if there exists a sequence  $\{Q_n\}$  of cross-cuts defining an end  $V$  such that  $\{Q_n\}$  converges to a point in  $\partial W$ , then  $V$  is prime (see, for example, [M1]).

Let  $\{V_n\}$  be a representative chain in a prime end  $V$ . Since each  $V_n$  is connected and  $\overline{W}$  is compact in  $Z$ ,  $\bigcap_{n \in \mathbb{N}} \overline{V_n}$  is a connected, compact, non-empty subset of  $Z$ . Thus it is either a single point or a continuum. We call  $I(V) = \bigcap_{n \in \mathbb{N}} \overline{V_n}$  the impression of the end  $V$ . The

impression of  $V$  is independent of the defining chain in  $V$ . (However, two prime ends can have the same impression. In Fig. 1, it appears that there are two prime ends corresponding to non-trivial circle points and that both have impressions that equal  $\partial W$ .) In [C], Carathéodory presents an example of a domain for which the impression of each prime end is a continuum; i.e., none is a single point. A point  $p$  in  $I(V)$  is called a principal point of  $V$  if there exists a sequence  $\{Q_n\}$  of crosscuts (defining a chain in  $V$ ) such that  $\{Q_n\}$  converges to  $p$ , i.e.,  $p$  is the *only* limit point of this sequence. The set of all such points is called the principal set of  $V$ . Finally, we say a point  $r$  in  $\partial W$  is accessible from  $W$  if there is an embedding  $\eta$  of  $(0,1]$  into  $W$  such that  $\lim_{t \rightarrow 0^+} \eta(t) = p$ . In Fig. 6, we illustrate these definitions. The following lemmas appear, for example, in [M1] (as Theorem 17.1 and Corollary 15, resp.):

LEMMA 5.1. The principal set of  $V$  has only one point  $e$  if and only if  $e$  is accessible from  $W$ .

LEMMA 5.2. The principal set of  $V$  is compact, connected, and non-empty.

Now we describe a topology on the set of prime ends. Let  $\mathcal{U}$  be an open set in  $W$ . We say an end  $V$  is contained in  $\mathcal{U}$  (i.e.,  $V \in \mathcal{U}$ ) if there exists a chain  $\{V_n\}$  in  $V$  all of whose elements are subsets of  $\mathcal{U}$ . Let  $W^* = W \cup \mathcal{E}$ , where  $\mathcal{E}$  is the set of prime ends of  $W$ . A set  $\mathcal{U}^*$  in  $W^*$  is open if and only if  $\mathcal{U}^* \cap W$  is open (in  $W$ ) and

$U^* \cap \mathcal{E} = \{V: V \text{ contains a chain all of whose elements lie in } W\}$ . With this topology, a sequence  $\{E_n\}$  of prime ends "converges" to a prime end  $E$ , represented by  $\{V_n\}$ , if for every  $m$ , there exists  $N(m) \in \mathbb{N}$  such that  $E_n \subseteq V_m$  for every  $n > N(m)$ . We call  $W^*$  together with this topology the prime end compactification of  $W$ . Central to the theory of boundary sets is the following theorem of Carathéodory (see, for example, [C-Lo]):

**THEOREM 5.3 (Carathéodory).** Let  $W$  be a connected, simply connected open set. Assume that  $\partial W$  has more than one point. Then  $W^*$  is homeomorphic to a closed disk, where points in  $W$  correspond to points in the interior of the disk, and the prime ends correspond to points in  $S^1$ , the boundary of the disk. Furthermore, if  $F$  is a continuous map on  $Z$  with  $W$  invariant under  $F$ , then there is map  $F^*$  on  $W^*$  so that  $F^* = F$  on  $W$ .

With this theorem, we are able to learn about the dynamics of  $F$  on the boundary of  $W$  by studying the corresponding dynamics on  $S^1$ , the boundary of  $D$ . Prime ends "map" to prime ends under  $F$ ; hence  $F$  induces a map  $F^*$  on  $W^*$ . Let  $\tau$  be a homeomorphism from  $W^*$  to  $\bar{D}$ , the closure of  $D$ . Then the circle  $S^1$  is invariant under the induced homeomorphism  $f = \tau \circ F^* \circ \tau^{-1}$  of  $\bar{D}$ . The study of homeomorphisms of the circle is classical. Here we mention briefly some facts about these maps which are needed in the arguments that follow. A reference for this material is [D].

Poincaré showed that associated with each orientation preserving homeomorphism  $\gamma$  of the circle is a "rotation" number, an asymptotic measure of the rotation of points on the circle under iteration by  $\gamma$ . In order to define this number, it is convenient first to consider a "lift" of  $\gamma$ . A map  $G$  of  $\mathbb{R}$  is called a lift of  $\gamma$  if  $\pi \circ G = \gamma \circ \pi$ , where  $\pi$  is the covering map from  $\mathbb{R}$  to  $S^1$ ; i.e.,  $\pi(x) = \exp(2\pi ix)$ . Let

$$p_G(x) = \lim_{n \rightarrow \infty} G^n(y)/n,$$

for  $x$  in  $S^1$  and  $y$  in  $\mathbb{R}$  such that  $\pi(y) = x$ . (The value of  $p_G(x)$  is independent of the choice of  $y$ .) We define the rotation number  $r$  of  $\gamma$  to be the unique number in  $[0,1)$  such that  $p_G(x) - r$  is an integer. This value is well-defined:

FACT C1. The value  $r = r(\gamma)$  is independent of both  $x$  and the particular lift  $G$  of  $\gamma$ .

The dynamics of  $\gamma$  are, to a large part, described by the rotation number  $r(\gamma)$ :

FACT C2. A map  $\gamma$  of the circle has points of *minimum* period  $q$  if and only if  $r(\gamma)$  is an irreducible fraction of the form  $p/q$ , for some positive integer  $p$ . The map  $\gamma$  has fixed points if and only if  $r(\gamma) = 0$ .

Thus, if  $\gamma$  has periodic points, they must all have the same period.

FACT C3. If  $\gamma$  has a periodic point of period  $n$ , then every point on the circle is either a fixed point of  $\gamma^n$  or is asymptotic to a fixed point under iterates of  $\gamma^n$ .

In the following set of definitions, we describe various notions of stability for periodic points and periodic prime ends. We often mention only fixed points, but the definitions and lemmas which follow carry over to periodic points by considering the appropriate iterate of  $f$ : a periodic point of period  $n$  is a fixed point of  $f^n$ .

A fixed point  $p$  on  $S^1$  is called:

(1) attracting on one side if all nearby points on one side of  $p$  converge to it under iteration by  $f$ ;

(2) repelling on one side if all nearby points on one side of  $p$  converge to it under iteration by  $f^{-1}$ .

The analogous definitions hold on the space of prime ends if the word "point" is replaced by the term "prime end", and if " $f$ " is replaced by " $F^*$ ".) By Fact C3, an isolated fixed point  $p$  on  $S^1$  is either attracting or repelling on each side. If  $p$  is attracting (resp., repelling) on one side, then by Carathéodory's Theorem, the associated prime end  $\mathcal{P}$  is attracting (resp., repelling) on one side.

A prime end  $\mathcal{P}$  fixed under  $F^*$  is called weakly stable from  $W$  if  $\mathcal{P}$  contains a chain  $\{V_n\}$  such that  $F(\overline{V_1}) \subseteq \overline{V_1}$ , for every  $i$ . The

following lemma follows easily from the definition of  $\partial W$  being unstable in  $W$  (see Sec. 1):

LEMMA 5.4. If  $\partial W$  is unstable in  $W$ , then no fixed prime end is weakly stable from  $W$ .

The following three lemmas are important in relating fixed points of  $F$  on  $\partial W$  to fixed points of  $f$  on  $S^1$ . Although there is a fixed prime end corresponding to each fixed point on the circle, it is not the case that a prime end which is fixed under  $F^*$  necessarily contains a point which is a fixed point of  $F$ . Lemma 5.5 appears in [C-L1].

LEMMA 5.5 (Cartwright-Littlewood). Let  $\mathcal{P}$  be a fixed prime end of  $F^*$ , and let  $\{Q_i\}$  be a chain of cross cuts converging to a point  $q$  (necessarily a principal point) of  $\mathcal{P}$ . If, for every  $i$ ,  $F(Q_i)$  has at least one point in common with  $Q_i$ , then  $q$  is a fixed point of  $F$ .

LEMMA 5.6. If  $\partial W$  is unstable in  $W$  and if a fixed prime end  $\mathcal{P}$  is attracting on one side, then all principal points of  $\mathcal{P}$  are fixed under  $F$ .

*Proof.* Suppose  $\mathcal{P}$  is attracting on one side. Let  $z$  be a principal point of  $\mathcal{P}$ . By Lemma 5.2, there exists a sequence  $\{Q_n\}$  of

cross-cuts converging to  $z$ . Let  $\{V_n\}$  be the chain defined by these crosscuts. By throwing out elements of the chain where necessary, we can assume that either  $F(Q_i) \cap Q_i \neq \emptyset$ , for all  $i$ , or that  $F(Q_i)$  is disjoint from  $Q_i$ , for all  $i$ . In the former case,  $z$  is fixed, by Lemma 5.5. Suppose that  $F(Q_i)$  is disjoint from  $Q_i$ , for all  $i$ . Then  $(\tau \circ F)(Q_i)$  is disjoint from  $\tau(Q_i)$ , for all  $i$ , and  $\tau(\mathcal{P}) = p$  is attracting on one side. Let  $\sigma_i$  on  $S^1$  be the end point of  $\tau(Q_i)$  which is on that side of  $p$ . Then for  $i$  sufficiently large,  $f^n(\sigma_i) \rightarrow p$ , as  $n \rightarrow \infty$ . Since  $\tau(Q_i)$  and  $(\tau \circ F)(Q_i)$  are disjoint, we then have that  $(\tau \circ F)(Q_i) \subset \tau(\overline{V_i})$ . But then  $F(\overline{V_i}) \subset \overline{V_i}$ , for all  $i$ , contradicting Lemma 5.4. Thus  $z$  is fixed under  $F$ . ■

*Proof of Theorem 1.1 (Attracting Lemma).* Suppose that  $x$  is a periodic circle point of period  $n$  and that  $x$  is attracting on one side. Then the corresponding prime end  $\mathcal{P}$  is fixed under  $(F^*)^n$  and attracting on one side. By Lemma 5.6 all principal points of  $\mathcal{P}$  are fixed under  $F^n$ . By Lemma 5.2 the set of principal points is connected. Since fixed points of  $F^n$  are isolated, there can be only one principal point, say  $p$ . By Lemma 5.1 the point  $p$  is accessible. For a given curve  $\Gamma$  in  $W$  limiting on  $p$ , the corresponding curve  $h^{-1}(\Gamma)$  (by definition) limits on a trivial point  $r$  in  $S^1$ .

## 6. Hyperbolicity

In this section we describe the dynamics on the set of accessible points under the hypotheses that  $F$  is a diffeomorphism of either the plane or the sphere and that periodic points in the boundary are hyperbolic. In addition, we either assume that  $W$  is a basin of attraction (see Sec. 1 for definition) or we add a condition on the map  $F$  at  $\infty$ . We say that  $\infty$  is repelling in  $\bar{W}$  if, for each  $r_1 > 0$ , there exists  $r_2 > 0$  such that if  $|x| < r_1$ , then  $|F^n(x)| < r_2$ , for all  $x$  in  $\bar{W}$  and  $n \geq 0$ .

**THEOREM 6.1.** Assume that the periodic points of  $F$  in  $\partial W$  are hyperbolic, and that either (i)  $W$  is a basin of attraction, or (ii)  $\partial W$  is unstable in  $W$  and  $\infty$  is repelling in  $\bar{W}$ . If the rotation number  $\rho$  is rational, then every accessible point either is a periodic point or is in the stable manifold of an accessible periodic point.

The following lemmas are used in the proof of Theorem 6.1. For each, the hypotheses of Theorem 6.1 are assumed. Let  $S$  be a (finite) periodic saddle of  $F$  in  $\partial W$ , and let  $W^S$  (resp.,  $W^u$ ) represent either branch of the stable (resp., unstable) manifold of  $S$ , excluding  $S$ .

**LEMMA 6.2.** If  $\partial W$  intersects  $W^S$ , then  $W^S$  and  $W$  are disjoint.

*Proof.* If  $W$  is a basin of attraction, then clearly  $W^S$  and  $W$  are disjoint. Suppose therefore that  $\partial W$  is unstable in  $W$ , that  $\infty$  is repelling in  $\bar{W}$ , and that both  $\partial W$  and  $W$  intersect  $W^S$ . Let  $Q_1$  be a crosscut in  $W \cap W^S$ , and let  $Q_2 = F(Q_1)$ . Then  $W - \{Q_1 \cup Q_2\}$  has three components. One component meets both  $Q_1$  and  $Q_2$ . Let  $D_1$  be the component that meets only  $Q_1$ , and let  $D_2$  be the component that meets only  $Q_2$ . Then  $D_2 = F(D_1)$ .

Since  $\infty$  is repelling in  $W$  and  $\bar{W}$  is invariant under  $F$ , there exists a compact set  $K$  such that  $F(K \cap \bar{W})$  is contained in  $K \cap \bar{W}$  and an open neighborhood of  $S$  is in  $K$ . Iterating  $D_1$  forward, there exists a sequence  $\{D_n\}$  of open sets in  $W$  intersecting  $W^S$  such that  $\{D_n\}$  approaches  $W^u$  (locally), as  $n \rightarrow \infty$ . Given  $\varepsilon > 0$ , choose  $j$  sufficiently large so that  $D_n$  intersects  $K$  and there is no  $\varepsilon$ -disk in  $D_n \cap K$  for all  $n > j$ . (This is possible since  $K$  includes an open neighborhood of  $S$  and there are only a finite number of  $\varepsilon$ -disks inside  $K$ .) Then for  $n > j$ , every point in  $D_n \cap K$  is within  $\varepsilon$  of the boundary, contradicting the hypothesis that  $\partial W$  is unstable in  $W$ . ■

LEMMA 6.3. If  $p \in S^1$  is a trivial fixed point, then it corresponds to an accessible fixed point  $S$  in the boundary  $\partial W$ . If  $S$  is a repeller, then so is  $p$ .

*Proof.* Corresponding to  $p$  is an accessible point  $S$  in  $\partial W$ . The point  $S$  is necessarily a fixed point since accessible points map to accessible points and  $S$  is the only accessible point corresponding to the prime end  $p$ .

Suppose that  $S$  is a repeller. Since the boundary  $\partial W$  is connected and more than one point, each circle centered at  $S$  of sufficiently small radius must intersect  $\partial W$ . Let  $\gamma$  be an "accessing" path in  $W$  which limits on  $S$  (corresponding to a path in the disk which limits on  $p$ ), and let  $\{Q_n\}$  be a sequence of crosscuts converging to  $S$  such that (1)  $Q_n$  is an arc of a circle of radius  $1/(n+N)$  for some fixed integer  $N \geq 1$ , and (2)  $\gamma$  intersects  $Q_n$  an odd number of times, for each  $n$ . As described in Sec. 5, since the sequence  $\{Q_n\}$  converges to one point (i.e., the point  $S$ ), it defines a prime end. Since this prime end has accessible point  $S$  with accessing path  $\gamma$ , it is represented by  $p$  on  $S^1$ . By the construction,  $p$  is a repeller on  $S^1$ . ■

We say that two accessing paths  $\gamma_0$  and  $\gamma_1$  are equivalent if  $\gamma_0$  can be homotoped to  $\gamma_1$  via a continuous family of paths that remains in  $W$ , all having the same endpoint  $S$ , (i.e., if there exists a continuous family  $g_t: I \rightarrow W$  such that  $g_0(I) = \gamma_0$ ,  $g_1(I) = \gamma_1$ , and  $g_t(0) = S$ , for all  $t \in I$ ). Notice that if  $S$  has two non-equivalent accessing paths, then it corresponds to (at least) two different circle points under  $h_c$ .

In the next two lemmas, we assume the following additional

hypotheses: (1)  $S$  is an accessible fixed point saddle; and (2)  $S$  has an associated trivial circle  $p$  which is attracting on at least one side, (i.e., there exists a point  $z \in S^1$ ,  $z \neq p$ , such that  $\lim_{n \rightarrow \infty} f^n(z) = p$ ).

For  $\varepsilon$  small, let  $M_\varepsilon$  be the union of the segments of the stable and unstable manifolds that connect  $S$  to the boundary of  $B_\varepsilon(S)$ , the  $\varepsilon$ -ball around  $S$ . We can assume that  $\varepsilon$  is small enough that the segments of the stable and unstable manifolds in  $M_\varepsilon$  intersect only at  $S$ .

LEMMA 6.4. Let  $\gamma$  be an accessing curve to  $S$ . Then  $\gamma$  is equivalent to an accessing curve that does not intersect  $M_\varepsilon$ .

*Proof.* Suppose that  $\gamma$  is not equivalent to an accessing curve that does not intersect  $M_\varepsilon$ . Since  $W$  is open, it must be the case that  $\gamma$  intersects at least two components of  $B_\varepsilon(S) - M_\varepsilon$  and that both  $\gamma$  and the boundary  $\partial W$  intersect  $W^S \cap M_\varepsilon$  or both intersect  $W^u \cap M_\varepsilon$ . The case in which both intersect  $W^S$  is ruled out by Lemma 6.2. Suppose that both intersect  $W^u$ . Let  $\{Q_n\}$  be a sequence of crosscuts converging to  $S$  such that  $Q_n$  is a closed interval on  $W^u$  and  $Q_n$  intersects  $\gamma$  an odd number of times, for each  $n$ . (Since the endpoints of  $Q_n$  are the only points of  $Q_n$  on the boundary  $\partial W$ , we can assume in fact that  $Q_n$  intersects  $\gamma$  only once.) The prime end determined by  $\{Q_n\}$  is represented by  $p$  on the circle. In this case,  $p$  must be a

repeller, a contradiction. ■

In the following, let  $\varepsilon > 0$  and let  $\gamma$  be an accessing path to  $S$  such that there is a unique component of  $B_\varepsilon(S) - M_\varepsilon$  that intersects  $\gamma$ . Call this component  $Q_\varepsilon$ . (The existence of  $Q_\varepsilon$  is guaranteed by Lemma 6.4.) Since  $S$  is hyperbolic, we can further assume that  $B_\varepsilon(S)$  is a neighborhood in which  $F$  is smoothly conjugate to a linear map, that  $S$  is the origin, and that  $Q_\varepsilon$  is an (open) quadrant in  $\mathbb{R}^2$ .

LEMMA 6.5. The component  $Q_\varepsilon$ , as defined above, contains no points of the boundary  $\partial W$ .

*Proof.* Suppose that  $Q_\varepsilon$  contains a point of  $\partial W$ . Let  $e_a$ ,  $a \in \mathbb{R}$ , be a family of ("hyperbolic-like") invariant curves in  $Q_\varepsilon$ . Since the boundary is connected, there is a connected component of  $\partial W \cap Q_\varepsilon$  containing  $S$  and a point  $b_a$  of  $e_a$ , for  $e_a$  sufficiently close to  $S$ . Assume  $e_a$  is sufficiently close to  $S$  that  $\gamma$  extends from  $S$  to a point  $g_a$  on  $e_a$ . Assume  $g_a$  is below  $b_a$  on  $e_a$  (the argument is similar if it is above). Assume further that  $F(g_a)$  is above  $b_a$ . (Otherwise take a higher iterate.) Then  $F(g_a)$  is between  $b_a$  and  $F(b_a)$  on  $e_a$ .

Since  $\gamma$  and  $f(\gamma)$  are both accessing curves to  $S$  (and they correspond to curves in the disk limiting on the same circle point),  $g_a$  and  $F(g_a)$  can be joined by a curve contained entirely in  $W$  so that

the resulting loop  $\beta$  is null-homotopic in  $W$ . This is a contradiction since either  $b_a$  or  $F(b_a)$  is contained in  $\beta$ . ■

*Proof of Theorem 6.1.* We assume that the rotation number is 0. (If the rotation number is  $p/q$  with  $p \neq 0$ , then replace  $F$  by  $F^q$  in the proof.) Let  $x$  be an accessible point in  $\partial W$  which is not a fixed point. Corresponding to  $x$  is a trivial circle point  $z$ . By Lemma 6.3,  $z$  is not a fixed point. Then the forward orbit of  $z$  converges to a fixed point  $p$  on  $S^1$ . By the Attracting Lemma,  $p$  is a trivial circle point. Corresponding to  $p$  is an accessible point  $S$  in  $\partial W$ . By Lemma 6.3,  $S$  is a fixed point. Since either  $W$  is a basin of attraction or  $\partial W$  is unstable in  $W$ ,  $S$  cannot be an attractor, and again by Lemma 6.3,  $S$  is not a repeller. Thus  $S$  is a saddle, and the hypotheses of Lemmas 6.4 and 6.5 are satisfied by  $S$ , since  $p$  is attracting on one side.

By Lemmas 6.4 and 6.5, there is at least one component  $Q_\epsilon$  of  $B_\epsilon(S) - M_\epsilon$  which is in  $W$  and contains no boundary points. If there are boundary points in another component of  $B_\epsilon(S) - M_\epsilon$ , then they are in connected components of  $\partial W$  which intersect both invariant manifolds bounding that component. If exactly one component is free of boundary points and is in  $W$ , then there are accessible points on one branch  $W^S$  of the stable manifold and one branch  $W^u$  of the unstable manifold. By Lemma 6.2, each point on this branch of  $W^S$  is an accessible boundary point. Thus points on one branch of the stable manifold of  $S$  are in one-to-one correspondence with points of  $S^1$  on one side of  $p$ .

Let  $\beta_1$  and  $\beta_2$  refer to the segments on either side of  $p$

consisting of points on the circle between  $p$  and the closest fixed points on either side. (If  $p$  is the only fixed point, then  $\beta_1 = \beta_2$ .) Let  $\beta_1$  be the segment which corresponds to  $W^S$ . Necessarily,  $\beta_1$  is part of the stable set of  $p$ . Let  $\{Q_n\}$  be a sequence of crosscuts converging to  $S$  such that one endpoint of  $Q_n$  is on  $W^u$  and one is on  $W^S$ , for each  $n$ . Since accessible boundary points on  $W^u$  converge under  $F^{-1}$  to  $S$ , given a point  $y$  in  $W^u \cap \partial W$  (necessarily accessible) and  $n > 0$ , all but a finite number of points in the forward orbit of  $y$  under  $F^{-1}$  will be in  $\overline{V_n}$ , the closure of the domain determined by  $Q_n$  and  $S$ . In this case  $p$ , which corresponds to the prime end determined by  $\{Q_n\}$ , is repelling on  $\beta_2$ . Since the forward orbit of  $z$  converges to  $p$ ,  $z$  must be on  $\beta_1$ , and thus  $x$  is in the stable manifold of  $S$ .

The argument given in the previous paragraph holds in all cases in which a sequence  $\{Q_n\}$  of crosscuts in  $W$  converging to  $S$  (i.e., a sequence which defines the prime end represented by  $p$ ) has the property that one endpoint of  $Q_n$  is in  $W^u$  and one is in  $W^S$ , for all  $n \geq 0$ . The case in which there are exactly three components of  $B_\varepsilon(S) - M_\varepsilon$  in  $W$  which are free of boundary points also reduces to this case. If the crosscuts do not have this property, then there are necessarily exactly two or exactly four components in  $W$ . In these cases, both endpoints of a crosscut are in one or the other branch of the stable manifold of  $S$ . (Since the fixed point  $p$  is attracting from at least one side on  $S^1$ , the case in which only the unstable manifold of  $S$  intersects the boundary is ruled out by an argument similar to that in the proof of Lemma 6.3.) In this case,  $p$  is necessarily attracting on the circle, and points on both  $\beta_1$  and  $\beta_2$  are in

one-to-one correspondence with points in the stable manifold of  $S$ .

Thus  $x$  is in the stable manifold of  $S$ . ■

The following corollaries follow from the proof of Theorem 6.1. The first extends Theorem 1.1 (the Attracting Lemma) to all points of  $S^1$ , not just periodic points. The second shows that the map  $h_c$ , the accessible-point extension of the Riemann map  $h$  (described in Sec. 1), is continuous on stable manifolds of periodic points of  $S^1$  (up to and including the periodic point). For a trivial circle point  $r$ , we let  $\bar{r}$  denote the corresponding accessible point in  $\partial W$ . We assume the hypotheses of Theorem 6.1.

COROLLARY 6.6. Assume that  $\rho$  is rational. If a point  $r$  in  $S^1$  is not a periodic point, then  $r$  is a trivial circle point.

COROLLARY 6.7. Let  $p$  in  $S^1$  be a periodic point of  $f$ , and let  $\{r_n\}$  be a sequence of points in  $S^1$  converging to a point  $r$ . If  $r_n$  is in the stable manifold of  $p$ , for each  $n$ , then the corresponding sequence  $\{\bar{r}_n\}$  of accessible points in  $\partial W$  converges to  $\bar{r}$  in  $\partial W$ .

## 7. Minimum Periods of Accessible Periodic Orbits

Unfortunately, although a rational rotation number  $p/q$  implies that  $f$  has a periodic orbit of minimum period  $q$  on  $S^1$ , we cannot claim that  $F$  has a periodic point of minimum period  $q$ . See, for example, the boundary depicted in Fig. 7, where  $\rho(\partial W, F)$  is  $1/3$ , and  $F$  has an accessible fixed point on the boundary but no period three orbit.

Recall that  $h_c$  is the accessible-point extension of the Riemann map  $h$ . If the rotation number  $\rho$  of  $f$  is rational (say  $p/q$ ), but not 0, then trivial circle points which are periodic (necessarily of minimum period  $q$ ) can map by  $h_c$  to periodic points in the plane of smaller minimum period. This situation is illustrated in Fig. 7, where all points in one orbit on the circle coalesce to a fixed point on the sphere. Surprisingly, Cartwright and Littlewood [C-L1] showed that this type of example is the only possible one when accessible points coalesce:

**THEOREM 7.1 (Cartwright-Littlewood).** If  $\rho \neq 0$ , then  $\partial W$  contains at most one accessible fixed point.

It is easily seen that this theorem rules out coalescing to an orbit of minimum period strictly between 1 and  $q$ . Suppose that a trivial periodic orbit of minimum period  $q$  on the circle maps (under  $h_c$ ) to a periodic orbit of minimum period  $k$  on  $\partial W$ , where  $k \neq 1$  and  $k \neq q$ .

Then  $k = q/r$  for some divisor  $r$  of  $q$  ( $r \neq 1$ ), and  $F^k$  has  $k$  accessible fixed points on  $\partial W$ . But the rotation number of the induced map  $f^k$  on the circle is non-zero, contradicting the theorem.

The situation illustrated in Fig. 7 can be largely overcome by using Theorem 6.1 and assuming that the accessible periodic points are saddles.

**PROPOSITION 7.2.** Assume the hypotheses of Theorem 6.1. If the rotation number  $\rho \neq 0$  is the reduced fraction  $p/q$ , where  $q \neq 2$ , then every accessible periodic saddle in  $\partial W$  has minimum period  $q$ .

*Proof.* Suppose there exists an accessible orbit of period  $k$  on  $\partial W$ , where  $1 < k < q$ . Then  $F^k$  has at least  $k$  fixed points, but the rotation number of the induced circle map  $f^k$  is non-zero, contradicting Theorem 7.1. Hence we assume there is an accessible fixed point saddle  $z$  on  $\partial W$ . Given a path  $\Gamma$  in  $W$  limiting on  $z$ , let  $y$  be the trivial circle point which is the limit point of  $h_c^{-1}(\Gamma)$ . Either  $y$  is a periodic point of period  $q$ , or the forward orbit of  $y$  under  $f^q$  converges to a periodic point  $r$ . By Theorem 1.1,  $r$  is a trivial circle point. By Corollary 6.4, the trivial circle point  $r$  corresponds to the accessible point  $z$  (i.e.,  $h_c(r) = z$ ), as do each of the  $q$  points  $r = r_1, r_2, \dots, r_q$  in the orbit of  $r$ .

Let  $O$  be the center of the disk which  $S^1$  bounds and let  $\gamma_1, \dots, \gamma_q$  be line segments joining  $O$  to  $r_1, \dots, r_q$ , respectively. Then  $h(\gamma_1), \dots, h(\gamma_q)$  are paths in  $W$ , all of which limit on  $z$ . Let  $r_i$  and  $r_j$  be adjacent points on the circle. Since  $r_1, \dots, r_q$  represent distinct prime ends on  $S^1$ , the closed loop formed by  $h(\gamma_i)$ ,  $h(\gamma_j)$ , and  $z$  necessarily contains boundary points in its interior. These boundary points are connected to  $z$  within the loop. Therefore, by Lemma 6.5,  $q$  can be at most 4.

If  $q$  is 3 or 4, then at least one branch of the stable manifold of  $z$  is in  $\partial W$ , and  $r$  is necessarily attracting on at least one side of the circle under  $f^q$  --as is each of the  $q$  points in the orbit of  $z$ . Each of these stable sets must correspond to a branch of the stable manifold of  $z$ . On the other hand, there exists a path  $\Gamma$  in the disk connecting a point in the stable set of  $r_i$  to a point in the stable set of  $r_j$  which crosses one of the segments  $\gamma_i$  or  $\gamma_j$  exactly once and intersects none of the other segments. Hence, in  $W$ ,  $h(\Gamma)$  crosses  $h(\gamma_i)$  (or  $h(\gamma_j)$ ) exactly once and intersects none of the other "accessing" paths, a contradiction for  $q > 2$ . ■

For a map of the sphere, two types of degeneracies are possible when  $\rho = 1/2$ , even with the hypothesis that accessible orbits are hyperbolic saddles. These possibilities are illustrated in Fig. 8. In Fig. 8a,  $\rho = 1/2$  and there is an accessible fixed point saddle  $p$

on  $\partial W$ . In Fig. 8b,  $\partial W$  is a line segment. The basin  $W$  (the complement of  $B$ ) is simply connected on the sphere. In this case,  $\rho = 1/2$  and there is an accessible fixed point saddle  $p$  and an accessible saddle orbit  $\{r_1, r_2\}$  of period two.

The situation is greatly simplified when we look at homeomorphisms of the plane. We use the following converse of Theorem 1.2 for planar maps:

**PROPOSITION 7.3.** Let  $F$  be a homeomorphism of the plane  $\mathbb{R}^2$ . If there exists an accessible fixed point in  $\partial W$ , then  $\rho = 0$ .

*Proof.* Suppose  $\rho \neq 0$ . Let  $x$  be an accessible fixed point, and let  $p \in S^1$  be a corresponding trivial circle point. Since  $\rho \neq 0$ , we can choose  $N > 1$  such that the intervals  $[p, f(p)]$ ,  $[f(p), f^2(p)]$ , ...,  $[f^{N-1}(p), f^N(p)]$  cover  $S^1$ . By Property 2 of the extension  $h_c$  of the Riemann map,  $h_c(f^i(p)) = x$ , for  $i \geq 1$ .

Let  $\gamma$  and  $\delta$  be paths beginning at a point  $O$  in  $D$  and ending at  $p$  and  $f(p)$ , respectively. Let  $\Gamma$  be the closed loop formed by  $h(\gamma)$ ,  $h(\delta)$ ,  $h(O)$ , and  $x$ . Choose a preferred direction, clockwise or counterclockwise, so that the accessible boundary points corresponding to trivial circle points between  $p$  and  $f(p)$  are in  $\Gamma$ . Let  $G$  be  $\Gamma$  together with its interior. Since the accessible

points are dense in  $\partial W$ , the entire boundary is contained in the compact set  $K = \bigcup_{i=1}^N f^i(G)$ . The complement of  $D$  is contained entirely in  $W$  or entirely in  $\mathbb{R}^2 \setminus \bar{W}$ . The former case is ruled out since  $W$  is simply connected. But then  $h(\gamma) \subset W$  is in the boundary of  $K$ , a contradiction. ■

Now assume that  $\rho$  is the reduced fraction  $p/q$ . Assume further that there is an accessible periodic orbit of minimum period  $r$  in  $\partial W$  (in the plane). The iterate  $F^r$  induces the map  $f^r$  on  $S^1$ . Since  $F^r$  has a fixed point, the rotation number of  $f^r$  is 0, by Prop. 7.3. Thus all periodic points of  $f$  in  $S^1$  are fixed points of  $f^r$ , which implies that  $q$  divides  $r$ . The next proposition shows that  $q$  must equal  $r$ .

**PROPOSITION 7.4.** If  $\rho = 0$ , then every accessible periodic point in  $\partial W$  is a fixed point of  $F$ .

*Proof.* Assume there is an accessible periodic point  $x$  of period  $q$ ,  $q > 1$ . Let  $p$  be a trivial circle point corresponding to  $x$ . By Property 2 of the map  $h_c$ ,  $p$  is not a fixed point. Let  $O$  be the center of  $D$ , let  $\gamma_0$  be the line segment from  $O$  to  $p$ , and let  $\gamma_i$  be the line segment from  $O$  to  $f^i(p)$ , for each  $i$ ,

$1 \leq i \leq q+1$ . Then  $h_c(\gamma_0)$  and  $h_c(\gamma_q)$  both contain  $p$ , and they form a closed curve  $\Gamma$  which, except for  $p$ , is contained in  $W$ .

Since  $\rho = 0$ ,  $f(p)$  is between  $p$  and  $f^q(p)$  on one side of the circle and  $f^{q+1}(p)$  is between them on the other side. Thus one of  $h_c(\gamma_1)$  and  $h_c(\gamma_{q+1})$  is inside  $\Gamma$  and the other is outside.

By Property 2 of the map  $h_c$ ,

$$h_c(f(p)) = F(h_c(p)) = F(x) = F^{q+1}(x) = F^{q+1}(h_c(p)) = h_c(f^{q+1}(p)).$$

But only one of  $h_c(f(p))$  and  $h_c(f^{q+1}(p))$  is in  $\Gamma$ , a contradiction. ■

**COROLLARY 7.5.** Let  $f$  be a homeomorphism of the plane  $\mathbb{R}^2$ . If  $\rho \neq 0$  is the reduced fraction  $p/q$ , then every accessible periodic point in  $\partial W$  has minimum period  $q$ .

*Proof.* Suppose there is an accessible periodic point with period  $r$ . By the discussion following Prop. 7.3,  $q$  divides  $r$ . It follows from Prop. 7.4 that since the rotation number of  $f^q$  (on  $S^1$ ) is 0,  $q$  must equal  $r$ . ■

The final corollary puts together the previous results with the assumption of hyperbolicity to obtain a statement that does not mention the rotation number  $\rho$ :

COROLLARY 7.6. Assume the following set of hypotheses:

- (1)  $F$  is a diffeomorphism of the plane  $\mathbb{R}^2$ ;
- (2) the periodic points of  $F$  in  $\partial W$  are hyperbolic;
- (3) either (i)  $W$  is a basin of attraction; or  
           (ii)  $\partial W$  is unstable in  $W$ , and  $\infty$  is repelling in  $\bar{W}$ ;

and (4) either (i) there exists an accessible periodic point of minimum period  $q$ , or  
           (ii) there exists an accessible point which converges (under  $f^q$ ) to a periodic point of minimum period  $q$ .

Then every accessible point in  $\partial W$  either is a periodic point of minimum period  $q$  or is in the stable manifold of such a periodic point.

*Proof.* We need to prove that if an accessible point  $x$  converges under  $f^q$  to a periodic point  $z$ , then  $z$  is accessible. First we show that  $\rho$  is rational. Assume otherwise. (For ease of exposition, we assume that  $q$  is 1 and that  $z$  is a fixed point. Otherwise, replace  $F$  with  $F^q$ .) By hypothesis,  $x$  is on one branch  $W^S$  of the stable manifold of  $z$ . Let  $y$  be a point in  $W$ , and let  $g_0$  and  $g_1$  be paths from  $y$  to  $x$  and from  $y$  to  $F(x)$ , respectively.

By Lemma 6.2,  $W$  and  $W^S$  are disjoint. (This lemma does not depend on any assumption about  $\rho$ .) Therefore, there must be accessible

points in the region  $V$  bounded by  $g_0$ ,  $g_1$ , and the portion of  $W^S$  between  $x$  and  $F(x)$ . In fact, since  $\rho$  is irrational, there must be points in the orbit of the accessible point  $x$  in the region  $V$ . But then  $W^S$  must enter  $V$ . The only way  $W^S$  can enter  $V$  is through  $g_0$ ,  $g_1$ , or  $W^S$ , all of which are impossible. (In particular,  $g_0$  and  $g_1$  are in  $W$ , which by Lemma 6.2 does not intersect  $W^S$ .) Therefore,  $\rho$  is rational, and by Theorem 6.1,  $x$  is in the stable manifold of an accessible periodic point, namely  $z$ .

Since  $z$  has minimum period  $q$  and  $\rho$  is rational, by Cor. 7.5, every accessible periodic point has minimum period  $q$ ; the result follows from Theorem 6.1. ■

## References

- [AS] K. Alligood and T. Sauer, "Rotation numbers of periodic orbits in the Hénon map", *Commun. Math. Phys.* 120 (1988), 105-119.
- [ATY] K. Alligood, L. Tedeschini-Lalli, and J. Yorke, "Metamorphoses: sudden jumps in basin boundaries", preprint.
- [B] G.D. Birkhoff, "Sur quelques courbes fermées remarquables", *Bull. Soc. Math. France* 60 (1932), 1-26.
- [BS] N.P. Bhatia and G.P. Szegő, Stability Theory of Dynamical Systems, Springer-Verlag, Heidelberg, 1970.
- [C] C. Caratheodory, "Über die Begrenzung einfach zusammenhängender Gebiete", *Math. Ann.* 73 (1913), 323-370.
- [C-L1] M.L. Cartwright and J.E. Littlewood, "Some fixed point theorems", *Ann. Math.* 54 (1951), 1-37.
- [C-L2] \_\_\_\_\_, "On non-linear differential equations of the second order: I. The equation  $y'' - k(1-y^2)y' + y = b\lambda k \cos(\lambda t + \alpha)$ ,  $k$  large", *J. London Math. Soc.* 20 (1945), 180-189.
- [C-Lo] E.F. Collingwood and A.J. Lohwater, Theory of Cluster Sets, Cambridge Tracts in Mathematics and Mathematical Physics, No. 56, Cambridge Univ. Press, 1966.
- [D] R.L. Devaney, An Introduction to Chaotic Dynamical Systems, Benjamin/Cummings Publishing Co., Menlo Park, 1986.
- [Do] A. Dold, Lectures on Algebraic Topology, Springer-Verlag, Heidelberg, 1972.
- [GOYa] C. Grebogi, E. Ott, and J. Yorke, "Basin boundary metamorphoses: changes in accessible boundary orbits", *Physica* 24D (1987), 243-262.
- [GOYb] \_\_\_\_\_, "Critical exponent of chaotic transients in nonlinear dynamical systems", *Phys. Rev. Lett.* 57 (1986), 1284-1287.
- [HJ] S. Hammel and C. Jones, "Jumping stable manifolds for dissipative maps of the plane", *Physica* 35D (1989), 87-106.
- [Li] M. Levi, Qualitative Analysis of the Periodically Forced Relaxation Oscillations, Mem. AMS 214, 1981.
- [Ln] N. Levinson, "A second order differential equation with singular solutions", *Annals of Math.* 50, no. 1 (1947), 127-153.

- [M1] J. Mather, "Topological proofs of some purely topological consequences of Caratheodory's theory of prime ends", Th.M. Rassias, G.M. Rassias, eds., Selected Studies, North-Holland (1982), 225-255.
- [M2] J. Mather, "Area preserving twist homeomorphisms of the annulus", *Comment. Math. Helvetici* 54 (1979), 397-404.
- [M3] \_\_\_\_\_, "Invariant subsets for area-preserving homeomorphisms of surfaces", *Advances in Math. Suppl. Studies*, Vol. 7B (1981).
- [Y] J.A. Yorke, "Dynamics, a Program for IBM-PC Clones", 1987.  
(Available from J. A. Yorke.)

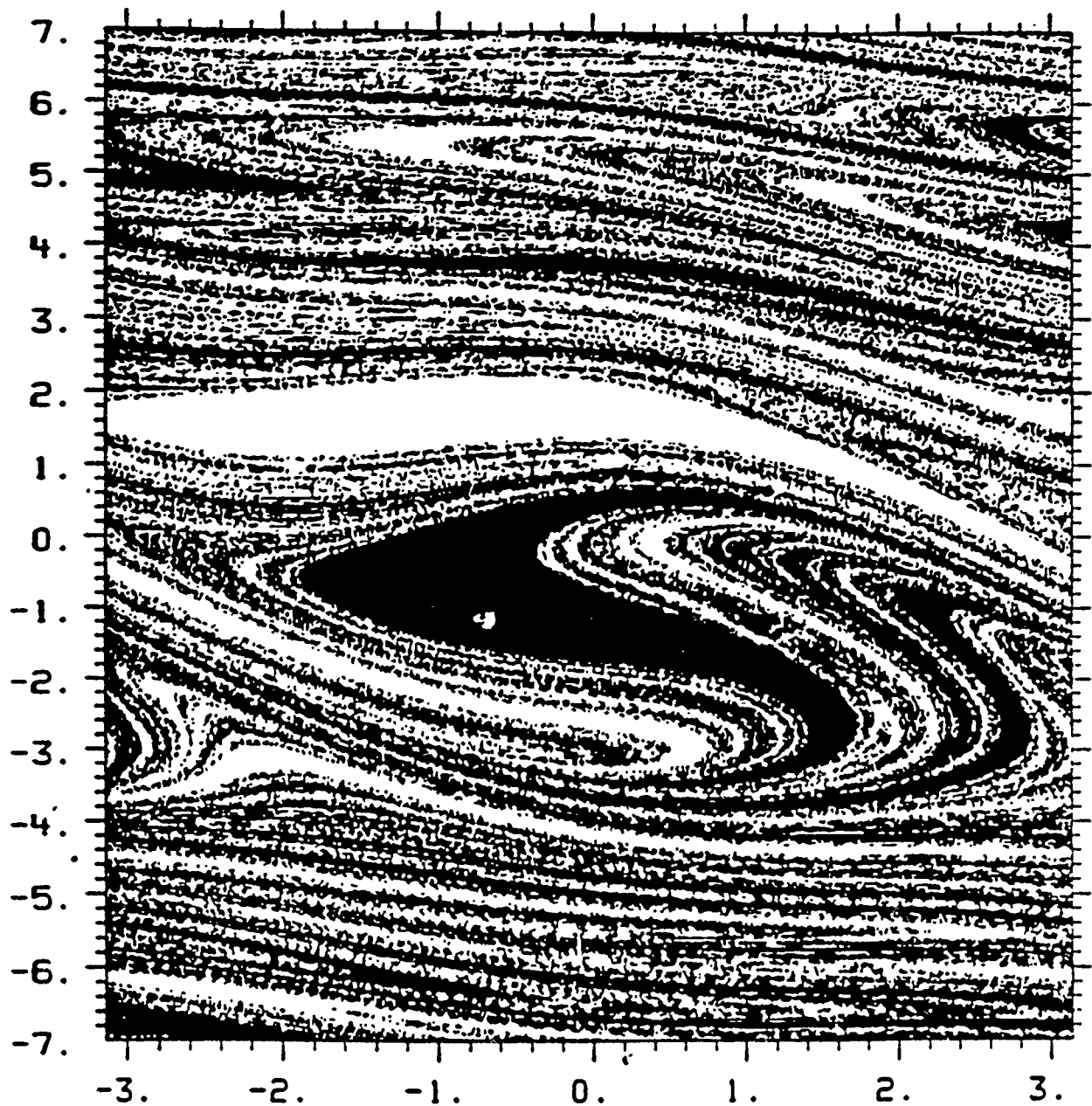


FIGURE 1. Two basins of attraction of the time  $2\pi$  map of the forced damped pendulum equation  $\theta'' + .1\theta' + \sin\theta = 2\cos t$  are shown in black and white. The black and white regions are connected on the cylinder.

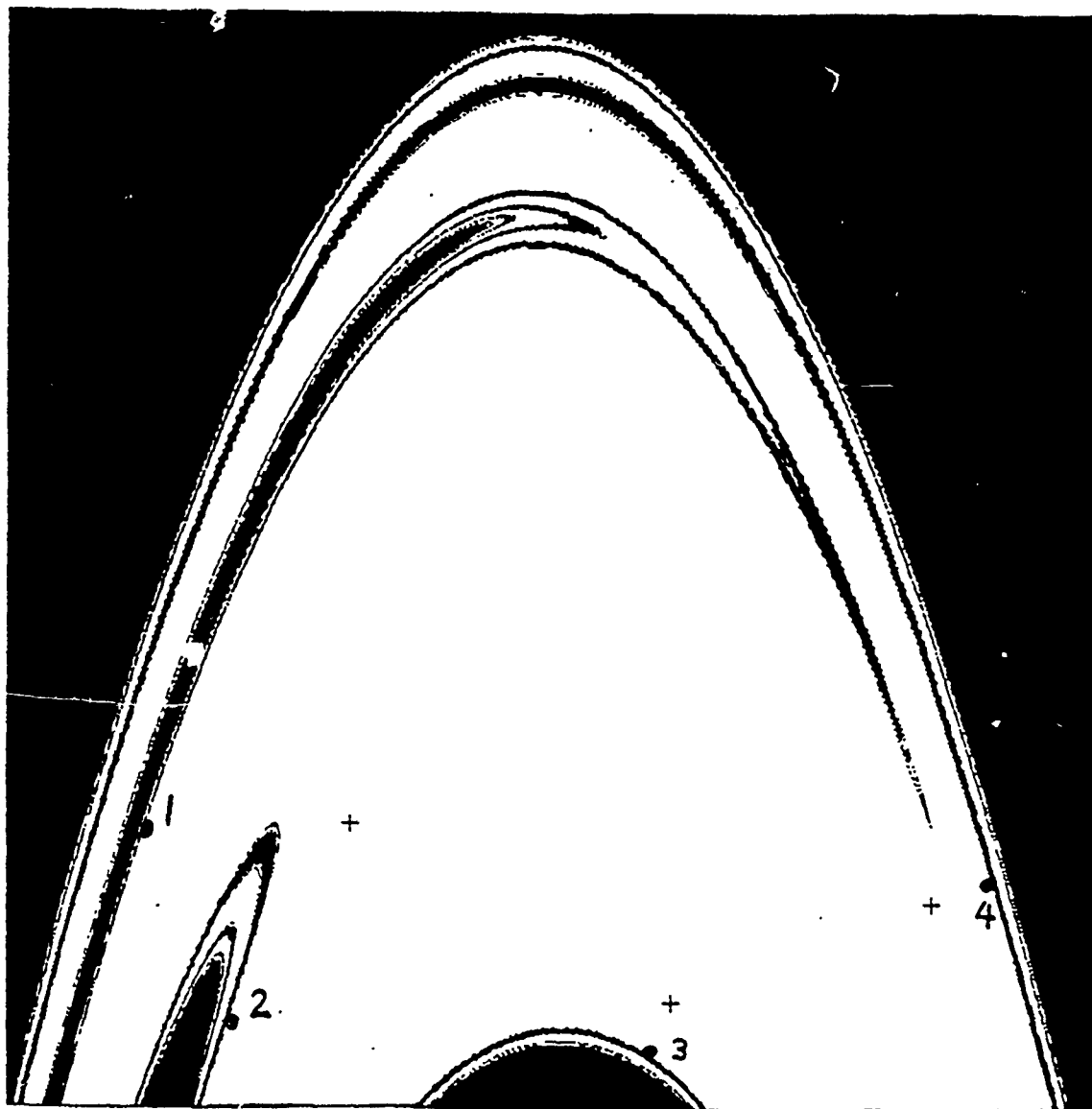


FIGURE 2. A portion of the basin of infinity of the Hénon map (3.1) is shown in black for  $b$  fixed at 0.3 and each of three values of the parameter  $\lambda$ . The  $x$  and  $y$  values shown are in the rectangle  $[-2, 2] \times [-2, 11]$ . In (a) at  $\lambda = 1.39$ , the set of accessible points consists of a period-four saddle and its stable manifold. Crosses show a period-three saddle to which the boundary jumps at a boundary metamorphosis at  $\lambda \approx 1.395$ . In (b) and (c) at  $\lambda = 1.40$  and  $\lambda = 1.42$ , respectively, the set of accessible points consists of this period-three saddle and its stable manifold.



FIG. 2(b)

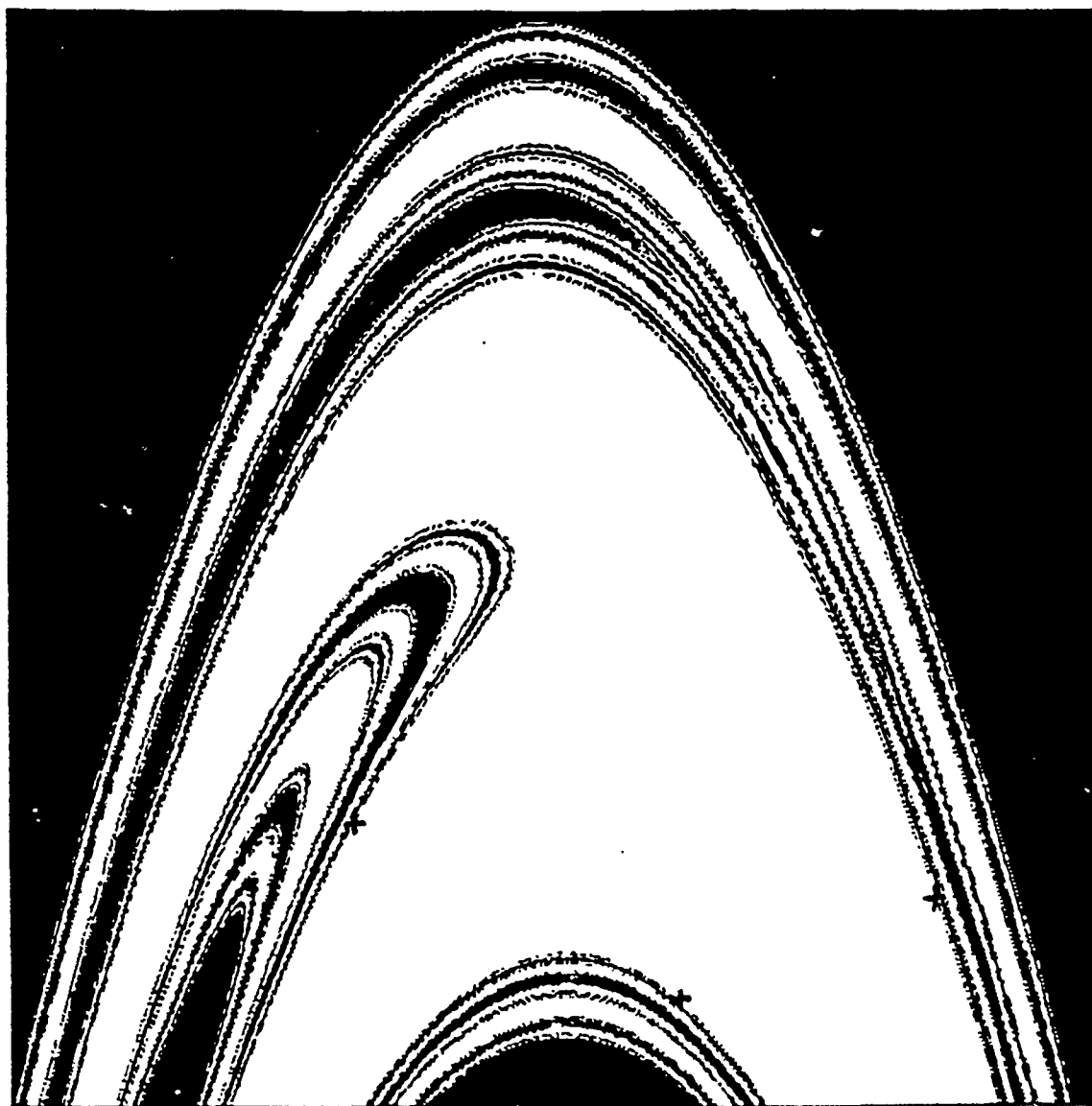


FIG. 2 (c)

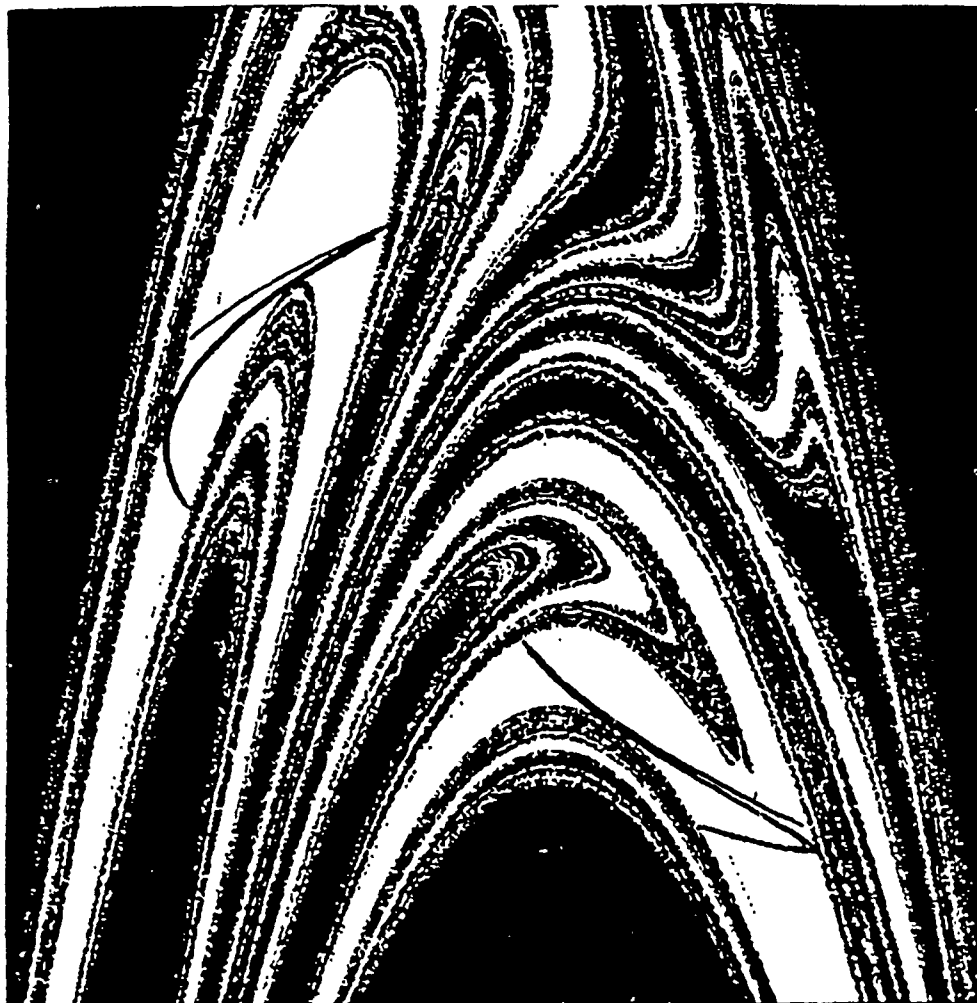


FIGURE 3. A portion of the basin of infinity of the orientation-reversing Hénon map (3.1) is shown in black. There is a two-piece attractor whose basin is not connected, and the basin boundary is fractal.

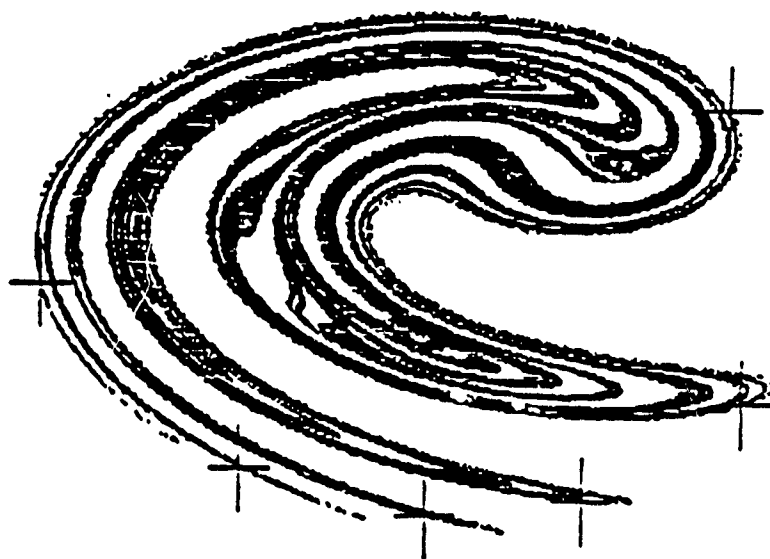


FIGURE 4. A chaotic attractor of the Ikeda map

$f(x,y) = (.97 + 0.9(x\cos\tau - y\sin\tau), 0.9(x\sin\tau + y\cos\tau))$ ,  
where  $\tau = 0.4 - 6.0/(1.0 + x^2 + y^2)$ , is shown. There is an  
accessible period 6 orbit on the attractor.

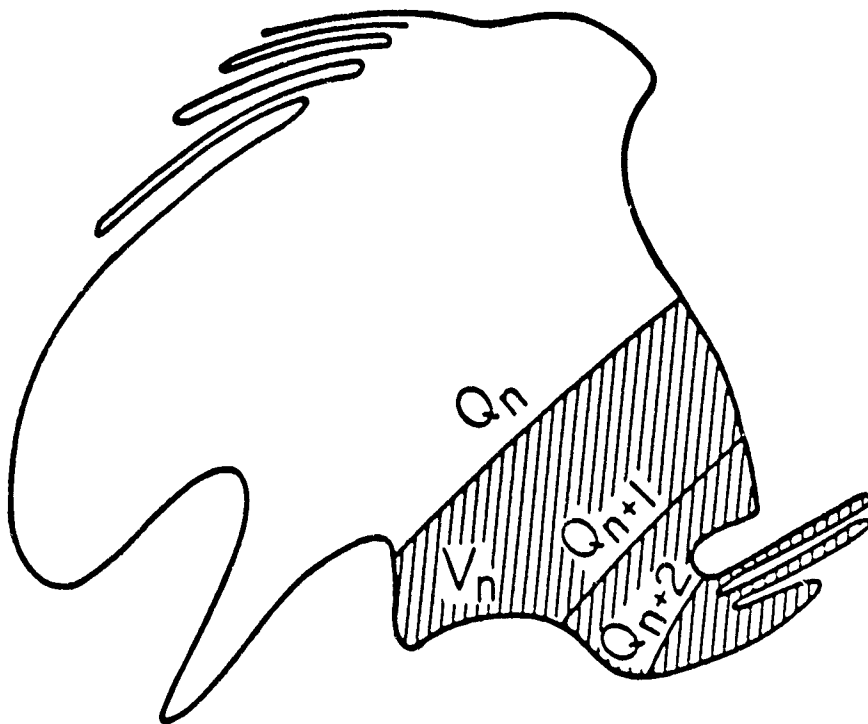
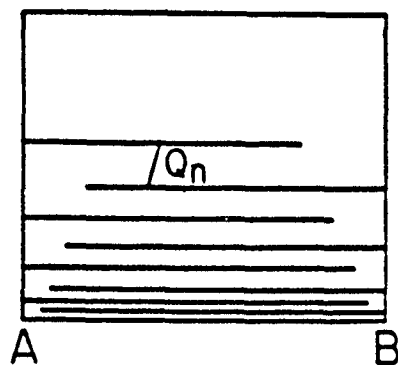
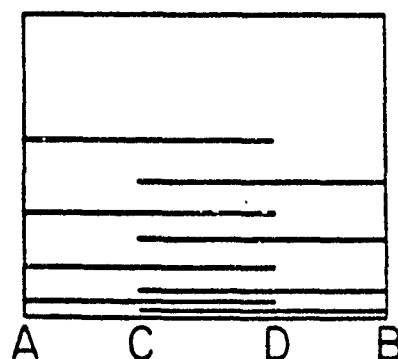


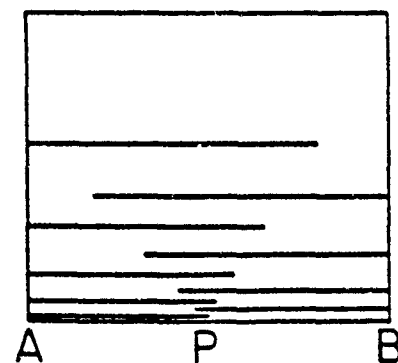
FIGURE 5. Sequences of crosscuts and subdomains defining a prime end are illustrates.



(a)



(b)



(c)

FIGURE 6. Each figure represents an open, simply connected set (the interior of the rectangle minus the line segments). In each case, segment AB is the impression of a prime end. In (a), each point of AB is a principal point, and there are no accessible points in AB. In (b), segment CD is the principal set of AB, and there are no accessible points in AB. In (c), P is the only principal point and the only accessible point of AB.

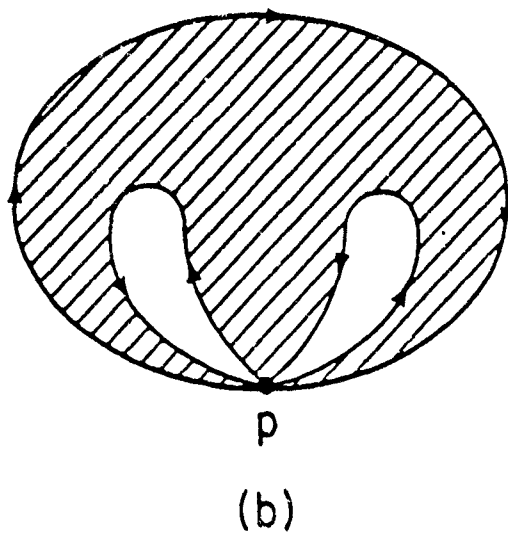
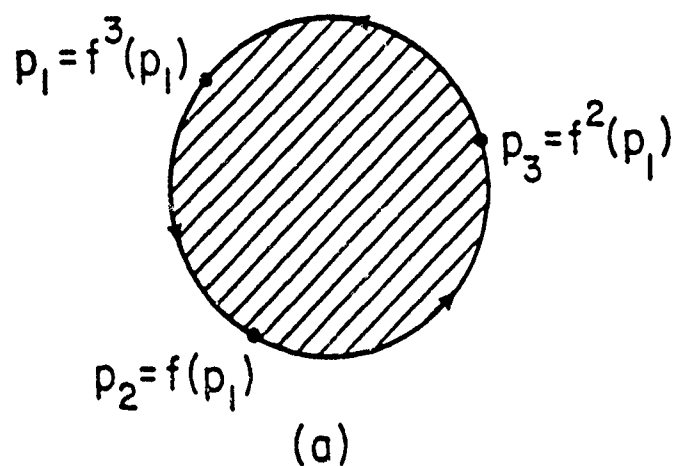
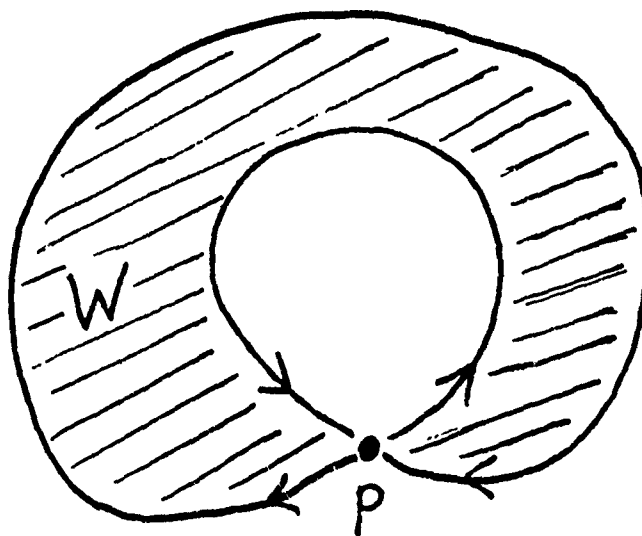


FIGURE 7. In (a) the rotation number on the boundary circle is  $1/3$ . The circle maps to the boundary in (b), however the boundary in (b) does not have an accessible periodic point of period 3, but rather has an accessible fixed point. This example is realizable on the sphere, not in the plane.

(a)



(b)

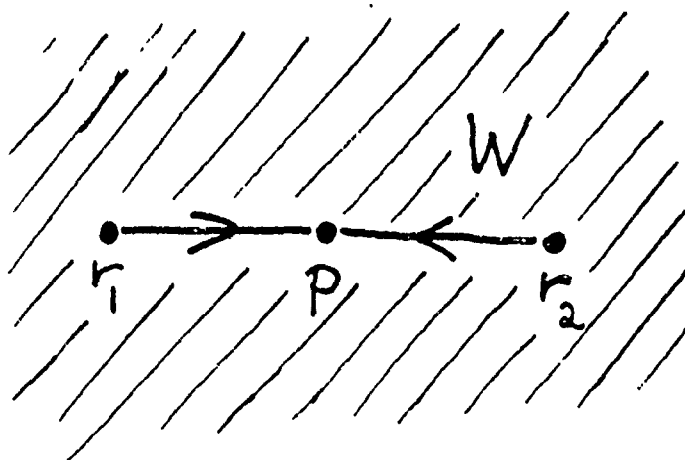


FIGURE 8. Two types of degeneracies on the sphere, in which the minimum periods of accessible saddles do not equal the periods of orbits of the associated circle map, are shown. By Prop. 7.2, the rotation numbers in each case is  $1/2$ . In (a), the boundary  $\partial W$  has a fixed point saddle  $p$ . In (b),  $W$  is the complement in the sphere of the line segment  $\partial W$  from  $r_1$  to  $r_2$ ; thus  $\partial W$  is the boundary of a simply connected set  $w$  on the sphere. Again, the rotation number is  $1/2$  and there is an accessible fixed point saddle  $p$  and an accessible saddle orbit  $\{r_1, r_2\}$  of period two.

Fig. 8

Figure 1

Two basins of attraction of the time  $2\pi$  map of the forced damped pendulum equation  $\Theta'' + .1\Theta' + \sin\Theta = 2\cos t$  are shown in black and white. The black and white regions are connected on the cylinder.

Figure 2

A portion of the basin of infinity of the Hénon map (3.1) is shown in black for  $b$  fixed at 0.3 and each of three values of the parameter  $\lambda$ . The  $x$  and  $y$  values shown are in the rectangle  $[-2,2] \times [-2,11]$ . In (a) at  $\lambda = 1.39$ , the set of accessible points consists of a period-four saddle and its stable manifold. Crosses show a period-three saddle to which the boundary jumps at a boundary metamorphosis at  $\lambda \approx 1.395$ . In (b) and (c) at  $\lambda = 1.40$  and  $\lambda = 1.42$ , respectively, the set of accessible points consists of this period-three saddle and its stable manifold.

Figure 3

A portion of the basin of infinity of the orientation-reversing Hénon map (3.1) is shown in black. There is a two-piece attractor whose basin is not connected, and the basin boundary is fractal.

Figure 4

A chaotic attractor of the Ikeda map

$$f(x,y) = (.97 + 0.9(x\cos\tau - y\sin\tau), 0.9(x\sin\tau + y\cos\tau)),$$

where  $\tau = 0.4 - 6.0/(1.0 + x^2 + y^2)$ , is shown. There is an accessible period 6 orbit on the attractor.

### Figure 5

Sequences of crosscuts and subdomains defining a prime end are illustrated.

### Figure 6

Each figure represents an open, simply connected set (the interior of the rectangle minus the line segments). In each case, segment AB is the impression of a prime end. In (a), each point of AB is a principal point, and there are no accessible points in AB. In (b), segment CD is the principal set of AB, and there are no accessible points in AB. In (c), P is the only principal point and the only accessible point of AB.

### Figure 7

In (a) the rotation number on the boundary circle is  $1/3$ . The circle maps to the boundary in (b), however the boundary in (b) does not have an accessible periodic point of period 3, but rather has an accessible fixed point. This example is realizable on the sphere, not in the plane.

### Figure 8

Two types of degeneracies on the sphere, in which the minimum periods of accessible saddles do not equal the periods of orbits of the associated circle map, are shown. By Prop. 7.2, the rotation number in each case is  $1/2$ . In (a), the boundary  $\partial W$  has a fixed point saddle p. In (b), W is the complement in the sphere of the line segment  $\partial W$  from  $r_1$  to  $r_2$ ; thus  $\partial W$  is the boundary of a simply connected set W on the sphere. Again, the rotation number is  $1/2$  and there is an accessible fixed point saddle p and an accessible saddle orbit  $\{r_1, r_2\}$  of period two.

# WHEN CANTOR SETS INTERSECT THICKLY

Erin R. Hunt<sup>1</sup>, Ittai Kan<sup>2</sup>, and James A. Yorke<sup>3</sup>

July 3, 1991

## Abstract

The thickness of a Cantor set on the real line is a measurement of its “size”. Thickness conditions have been used to guarantee that the intersection of two Cantor sets is nonempty. We present sharp conditions on the thickness of two Cantor sets which imply that their intersection contains a Cantor set of positive thickness.

## 1 Introduction

Newhouse defined [5] a nonnegative quantity called the “thickness” of a Cantor set in order to formulate conditions which will guarantee that two Cantor sets intersect. (All Cantor sets considered in this paper lie in  $\mathbb{R}^1$ .) These conditions have been used [5, 6, 7, 8, 9] in the study of two-dimensional dynamical systems to deduce the existence of tangencies between stable and unstable manifolds whose one-dimensional cross sections are Cantor sets.

Thickness may be thought of as a measure of how large a Cantor set is relative to the intervals in its complement. Henceforth, these intervals will be referred to as *gaps*; the two unbounded intervals in the complement are each included in our use of the term gap. Newhouse’s result [5, 7, 8] is that two Cantor sets must intersect if the product of their thicknesses is at least one, and neither set lies in a gap of the other. When this latter condition is satisfied, the sets are said to be *interleaved*. In [10], Williams observed the surprising fact that two interleaved Cantor sets can have thicknesses well above one and still only intersect in a single point. One might hope that under sufficiently strong

---

<sup>1</sup>Code R44, Naval Surface Warfare Center, Silver Spring, MD 20903-5000

<sup>2</sup>Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030

<sup>3</sup>Institute for Physical Science and Technology and Department of Mathematics, University of Maryland, College Park, MD 20742

The first author was supported by the ONT Postdoctoral Fellowship Program administered by ASEE, by ONR, and by the NSWC Independent Research Program. All three authors were partially supported by the Applied and Computational Mathematics Program of DARPA.

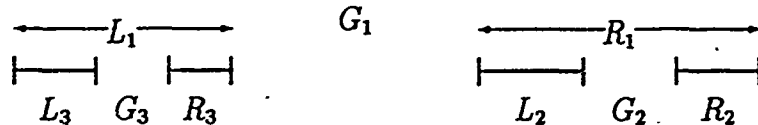


Figure 1: Constructing a Cantor set

thickness conditions, the intersection would be a Cantor set. However, the intersection of two arbitrarily thick interleaved Cantor sets can contain isolated points, so Williams posed the question of what conditions on the thicknesses of two interleaved Cantor sets will guarantee that their intersection contains another Cantor set. Williams obtained such a condition, though it is not sharp. In this paper we obtain the sharp condition. More precisely, we find a curve in  $(\tau_1, \tau_2)$ -space such that if the ordered pair  $(\tau_1, \tau_2)$  of thicknesses of two interleaved Cantor sets lies above the curve, their intersection contains a Cantor set, but if the pair of thicknesses lies below the curve there exist examples for which the intersection is a single point. Kraft [2] has independently arrived at this condition. We further show that if the thickness pair lies above the curve, the intersection must contain a Cantor set of positive thickness. This is the only result that addresses in terms of thickness how large the intersection of two Cantor sets must be. There are well known probabilistic results concerning the Hausdorff dimensions of intersections of Cantor sets (c.f. [1, 3, 4]).

One may think of a Cantor set as being constructed by starting with a closed interval and successively removing open gaps in order of decreasing length. Williams' formulation of the thickness of a Cantor set may then be thought of as follows. Each gap  $G_n$  is removed from a closed interval  $I_n$ , leaving behind closed intervals  $L_n$ , the left piece of  $I_n - G_n$ , and  $R_n$  on the right (see Figure 1.) Let  $\rho_n$  be the ratio of the length of the smaller of  $L_n$  and  $R_n$  to the length of  $G_n$ . The thickness of the set is the infimum of  $\rho_n$  over all  $n$ .

We consider as an example the "middle-third" Cantor set, constructed as follows. Start with the closed interval  $[0, 1]$ , and remove the open interval  $(1/3, 2/3)$ , the middle third of the original interval. Then from each of the two remaining intervals, remove their middle thirds; repeat this process infinitely often. Each gap  $G_n$  is the same length as the adjacent intervals  $L_n$  and  $R_n$ , so  $\rho_n = 1$  for all  $n$ . Thus the thickness of the middle-third Cantor set is one.

There is a connection between the thickness of a Cantor set and its fractal dimension, which depends in part on how the ratios  $\rho_n$  are distributed as  $n \rightarrow \infty$ . However, two large gaps close together make the thickness of a set very small, while its dimension can still be large. It was shown in [7] that the Hausdorff dimension of a Cantor set with thickness  $\tau$  is bounded below by  $\log 2 / \log(2 + 1/\tau)$ . This lower bound is sharp for the middle-third Cantor set (whose dimension is  $\log 2 / \log 3$ .)

We offer here a new formulation of the definition of thickness which we state for all

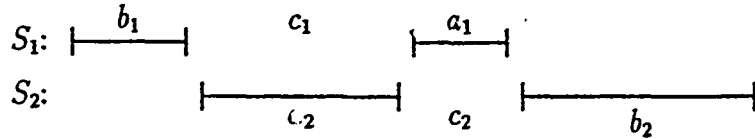


Figure 2: Non-intersecting interleaved sets

compact sets, not just Cantor sets. (The results in this and previous papers are found to be valid for all compact sets.) We define non-degenerate intervals to have infinite thickness, while singletons are defined to have thickness zero. In fact, any set containing an isolated point will be seen to have thickness zero. To define the thickness of a compact set  $S$  which is not an interval, we consider a type of subset of  $S$  obtained by intersecting  $S$  with a closed interval. We call such an intersection  $P$  a *chunk* of  $S$  if  $P$  is a proper subset of  $S$  and has a positive distance from  $S - P$ , the complement of  $P$  in  $S$ . (Notice that for  $P$  to be a chunk both  $P$  and  $S - P$  must be closed and nonempty.) We then define the thickness of  $S$  to be the infimum over all chunks  $P$  of the ratio between the diameter of  $P$  and the distance from  $P$  to  $S - P$ . In the case of the middle-third Cantor set, the given ratio can be shown to be smallest when the chunk  $P$  is obtained by intersecting  $S$  with an interval  $L_n$  or  $R_n$ , in which case the ratio is one. In Section 2 we will show that our new definition is equivalent to the old one for all Cantor sets.

The reason thickness is an appropriate quantity for determining when one can guarantee that two compact sets intersect is illustrated by considering an example where each of the two sets is a union of two disjoint intervals. For  $i = 1, 2$  let  $S_i$  consist of closed intervals of lengths  $a_i$  and  $b_i$  with  $a_i \leq b_i$ , separated by a distance  $c_i$ . Then each  $S_i$  has only two chunks, and is found to have thickness  $a_i/c_i$ . If the product of the thicknesses  $a_1 a_2 / c_1 c_2$  is at least one, then either  $a_1 \geq c_2$  or  $a_2 \geq c_1$  (or both); assume  $a_1 \geq c_2$ . Then since  $b_1 \geq a_1$ , neither interval of  $S_1$  can lie in the gap of  $S_2$ ; hence if the two sets are interleaved, they must intersect. If on the other hand  $a_1 a_2 / c_1 c_2 < 1$ , then with an affine map we can scale the sets so that  $a_1 < c_2$  and  $a_2 < c_1$ , and position them so that the component of  $S_1$  with length  $a_1$  lies inside the gap of  $S_2$ , and vice versa. The two sets are then interleaved, but they do not intersect (see Figure 2). This example could of course be made to involve Cantor sets by constructing very thick Cantor sets in each chunk of each  $S_i$ .

An important point which is apparent in the above example is that the union of two sets can have a smaller thickness than either of the original sets. In other words, adding points to a set can decrease its thickness. By the same token, one may be able to increase the thickness of a set by removing appropriate subsets. This observation is useful in the following way. No matter how thick two interleaved compact sets are, their intersection may have thickness zero because it may contain isolated points, or arbitrarily small chunks which are relatively isolated from the rest of the intersection. Nonetheless we are able to show that

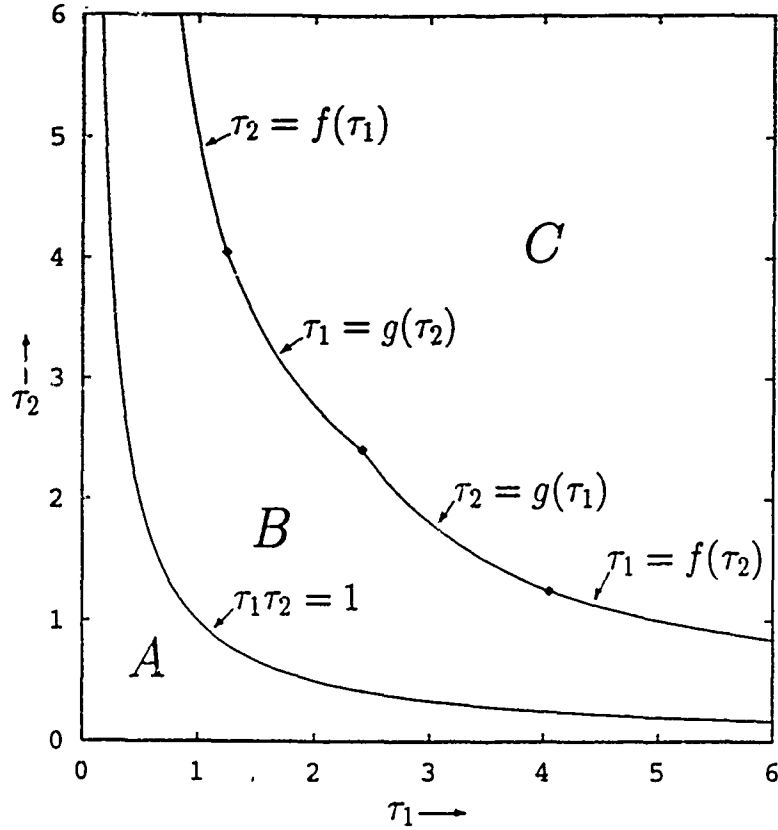


Figure 3: The intersection of two interleaved compact sets with thicknesses  $\tau_1$  and  $\tau_2$  can be empty for  $(\tau_1, \tau_2)$  in region  $A$ , must be nonempty but can be a single point in region  $B$ , and must contain a set of positive thickness in region  $C$ .

if the original sets are thick enough, then by throwing out the relatively isolated parts of their intersection we can obtain a set of positive thickness in the intersection.

To define the set  $C$  of thickness pairs  $(\tau_1, \tau_2)$  for which a Cantor set of intersection can be guaranteed, we make use of the functions

$$f(\tau) = \frac{\tau^2 + 3\tau + 1}{\tau^2},$$

$$g(\tau) = \frac{(2\tau + 1)^2}{\tau^3}.$$

Let  $C$  be the set of pairs  $(\tau_1, \tau_2)$  for which one of the following sets of conditions holds:

$$\tau_1 \geq \tau_2, \quad \tau_1 > f(\tau_2), \quad \text{and} \quad \tau_2 > g(\tau_1) \quad (1.1)$$

or

$$\tau_2 \geq \tau_1, \quad \tau_2 > f(\tau_1), \quad \text{and} \quad \tau_1 > g(\tau_2) \quad (1.2)$$

(see Figure 3.) Our main result is as follows.

**Theorem 1** *There is a function  $\varphi(\tau_1, \tau_2)$  which is positive in region  $C$  such that for all interleaved compact sets  $S_1, S_2 \subset \mathbb{R}$  with  $\tau(S_1) \geq \tau_1$  and  $\tau(S_2) \geq \tau_2$ , there is a set  $S \subset S_1 \cap S_2$  with thickness at least  $\varphi(\tau_1, \tau_2)$ .*

Notice that a compact set with positive thickness can have no isolated points, and thus must either be a Cantor set or contain an interval; either way it contains a Cantor set.

We remark that  $(\tau_1, \tau_2)$  is in  $C$  if both thicknesses are greater than  $\sqrt{2} + 1$ . This is the critical value Williams found for the case of interleaved Cantor sets with the same thickness. Also, no matter how small one thickness is, the other thickness can be chosen large enough so that the pair lies in  $C$ . Our results and the results of Newhouse are summarized in Figure 3.

In Section 2 we give a proof of Newhouse's result, which will illustrate some of the methods to be used later. Then we present for all pairs  $(\tau_1, \tau_2)$  not in  $C$  an example of interleaved compact sets with thicknesses  $\tau_1$  and  $\tau_2$  whose intersection is a single point (except when  $(\tau_1, \tau_2)$  is on the boundary of  $C$ , in which case our example gives a countable intersection.) This example shows that Theorem 1 is sharp in that its conclusion cannot hold for any larger set of thickness pairs  $(\tau_1, \tau_2)$ . In Section 3 we prove Theorem 1, and in Section 4 we discuss some further properties of  $S_1 \cap S_2$ . The positive thickness set  $S \in S_1 \cap S_2$  constructed in Section 3 need not be dense in  $S_1 \cap S_2$ ; however we find that there are subsets with thickness at least  $\varphi(\tau_1, \tau_2)$  arbitrarily near any accumulation point of  $S_1 \cap S_2$ . In addition, we find bounds on the diameter of  $S$  which allow us to obtain thickness conditions that imply that the intersection of three Cantor sets is nonempty.

## 2 Preliminaries

Let us define precisely the concepts and notation we will use.

**Definition 1** *We say two sets  $S_1, S_2 \subset \mathbb{R}$  are interleaved if each set intersects the interior of the convex hull of the other set (that is, neither set is contained in the closure of a gap of the other set.)*

We define the distance between two nonempty sets  $S_1, S_2$  to be

$$d(S_1, S_2) = \inf\{|x - y| \mid x \in S_1, y \in S_2\},$$

and write  $S_2 - S_1$  for the intersection of  $S_2$  with the complement of  $S_1$ . We say that a set  $S_1$  is a *chunk* of a set  $S_2$ , and write  $S_1 \propto S_2$ , if  $S_1$  is the intersection of a closed interval with  $S_2$ , is a proper subset of  $S_2$ , and  $d(S_1, S_2 - S_1) > 0$ . Notice that a closed set  $S$  has a chunk if and only if it is not connected. We denote the diameter of a set  $S$  (the length of its convex hull) by  $|S|$ .

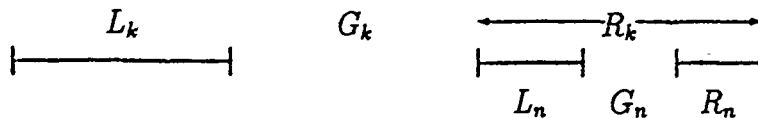


Figure 4: Chunks and gaps of a Cantor set ( $k < n$ )

**Definition 2** Given a compact set  $S \subset \mathbb{R}$ , we define the thickness of  $S$  to be

$$\tau(S) = \inf_{P \in S} \frac{|P|}{d(P, S - P)} \quad (2.1)$$

provided  $S$  has a chunk. Otherwise, we let  $\tau(S) = 0$  if  $S$  is empty or consists of a single point, and  $\tau(S) = \infty$  if  $S$  is an interval with positive length.

The following simple proposition demonstrates that Definition 2 agrees with Williams' definition of thickness for Cantor sets [10].

**Proposition 2** Let  $S$  be a Cantor set, and define the ratios  $\rho_n$  as in the introduction. Then the quantity  $\tau(S)$  given by (2.1) is equal to the infimum of  $\rho_n$  over all  $n$ .

**Proof** The intervals  $L_n$  and  $R_n$  defined in the introduction are the convex hulls of chunks  $A_n = L_n \cap S$  and  $B_n = R_n \cap S$  of  $S$ . Since the gap  $G_n$  is not larger than any previously removed gap  $G_k$ ,  $k < n$ , it follows that

$$d(A_n, S - A_n) = d(B_n, S - B_n) = |G_n|$$

(see Figure 4.) Thus for all  $n$ ,

$$\rho_n = \min \left( \frac{|L_n|}{|G_n|}, \frac{|R_n|}{|G_n|} \right) = \min \left( \frac{|A_n|}{d(A_n, S - A_n)}, \frac{|B_n|}{d(B_n, S - B_n)} \right) \geq \tau(S).$$

Next, if  $P$  is a chunk of  $S$ , it must be bordered on each side by a gap of  $S$ ; let  $G_n$  be the smaller of these two gaps. Then  $|G_n| = d(P, S - P)$  and  $|P| \geq \min(|L_n|, |R_n|)$ . Therefore

$$\tau(S) = \inf_{P \in S} \frac{|P|}{d(P, S - P)} \geq \inf_n \rho_n,$$

which completes the proof ■

We now prove Newhouse's result in a way that will motivate our later examples and methods.

**Proposition 3** If  $S_1$  and  $S_2$  are interleaved compact sets with  $\tau(S_1) \cdot \tau(S_2) \geq 1$ , then  $S_1 \cap S_2$  is not empty.

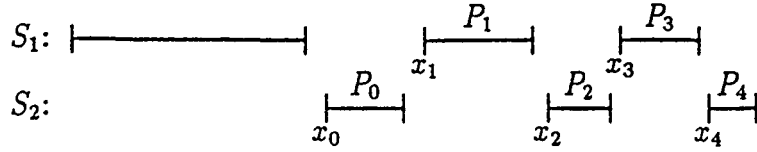


Figure 5: The points  $x_n$  and chunks  $P_n$

**Proof** Let  $S_1$  and  $S_2$  be as above, and let

$$x_0 = \max \left( \inf_{x \in S_1} x, \inf_{x \in S_2} x \right),$$

the greater of the leftmost points of  $S_1$  and  $S_2$ . Assume without loss of generality that  $x_0 \in S_2$ . We will show that  $S_1 \cap S_2$  is nonempty by looking for the leftmost point of this set. Let  $x_1$  be the leftmost point of  $S_1$  which is at least as great as  $x_0$ . Since  $S_1$  and  $S_2$  are interleaved,  $x_1$  must exist (otherwise  $S_1$  would lie entirely to the left of  $S_2$ ; see Figure 5.) Next, let  $x_2$  be the leftmost point of  $S_2$  greater than or equal to  $x_1$ . Once again the interleaving assumption implies that  $x_2$  exists, for otherwise  $S_2$  would lie inside a gap of  $S_1$ . We similarly define  $x_3, x_4, \dots$ ; if each of these points can be shown to exist, we claim to be done. Then  $\{x_n\}$  will be a nondecreasing sequence which is bounded above (since  $S_1$  and  $S_2$  are bounded), so it approaches a limit. This limit must belong to both  $S_1$  and  $S_2$  since these sets are closed and the odd numbered terms of  $\{x_n\}$  belong to  $S_1$ , the even ones to  $S_2$ .

If at any step  $x_n$  exists and equals  $x_{n-1}$ , then  $x_{n+1}, x_{n+2}, \dots$  will also equal  $x_{n-1}$ , and we will have found a point in  $S_1 \cap S_2$ . Henceforth we assume  $x_0 < x_1 < \dots$  as long as they are defined. We know at least that  $x_0, x_1$ , and  $x_2$  exist, so there is a chunk  $P_0$  of  $S_2$  which lies in  $[x_0, x_1]$ , whose diameter is thus less than  $x_1 - x_0$ , and whose distance from the rest of  $S_2$  is greater than  $x_2 - x_1$  (see again Figure 5.) Then

$$\frac{x_1 - x_0}{x_2 - x_1} > \frac{|P_0|}{d(P_0, S_2 - P_0)} \geq \tau(S_2). \quad (2.2)$$

Let  $P_1$  be the largest chunk of  $S_1$  which lies in  $[x_1, x_2]$ . If  $x_3$  did not exist, in other words if all points in  $S_1$  were less than  $x_2$ , then  $S_1 - P_1$  would lie to the left of  $P_1$ , and the distance between these sets would be greater than  $x_1 - x_0$ . But then using (2.2) and  $\tau(S_1) \cdot \tau(S_2) \geq 1$  we would have

$$\frac{|P_1|}{d(P_1, S_1 - P_1)} < \frac{x_2 - x_1}{x_1 - x_0} < \frac{1}{\tau(S_2)} \leq \tau(S_1),$$

contradicting the definition of the thickness of  $S_1$ . Thus  $x_3$  exists, and similarly to (2.2) we obtain

$$\frac{x_2 - x_1}{x_3 - x_2} > \frac{|P_1|}{d(P_1, S_1 - P_1)} \geq \tau(S_1). \quad (2.3)$$

Likewise (2.3) can be used to show the existence of  $x_4$ , and so forth. The proof is completed by induction. ■

One could similarly find the rightmost point in  $S_1 \cap S_2$ , but as Williams observed it may coincide with the leftmost point, even if both thicknesses are significantly greater than 1. We next present an example which will give a single point of intersection for thickness pairs  $(\tau_1, \tau_2)$  not in the closure of region  $C$ , and a countable intersection for  $(\tau_1, \tau_2)$  on the boundary of  $C$ . In our example both sets are countable unions of closed intervals, but they could be replaced by Cantor sets with the same thicknesses by constructing a very thick Cantor set in each of the closed intervals.

Let  $\tau$  be a positive constant, and define the intervals

$$A_0 = [\tau^2 + 3\tau + 1, (2\tau + 1)^2],$$

$$B_0 = [\tau^2, \tau^2 + 3\tau + 1],$$

$$A_n = \left(-\frac{\tau}{2\tau + 1}\right)^n A_0,$$

$$B_n = \left(-\frac{\tau}{2\tau + 1}\right)^n B_0,$$

where multiplication of a set by a scalar means the set obtained by multiplying each element of the original set by the given scalar. Let

$$S_1 = \left(\bigcup_{n=0}^{\infty} A_n\right) \cup \{0\}, \quad S_2 = \left(\bigcup_{n=0}^{\infty} B_n\right) \cup \{0\}.$$

Notice that  $B_n$  is the closure of the interval between  $A_n$  and  $A_{n+2}$  for all  $n$ , and  $A_n$  is the closure of the interval between  $B_{n-2}$  and  $B_n$  for  $n \geq 2$ . Thus  $S_1 \cap S_2$  is countable, containing only the point 0 and endpoints of the intervals  $A_n$  and  $B_n$ . Furthermore, the intersection could be reduced to only the point 0 by shrinking the intervals which make up one of the sets by a factor arbitrarily close to one.

Let us compute the thicknesses of the sets  $S_1$  and  $S_2$ . Observe that

$$|A_n| = d(B_{n-2}, B_n) = \left(\frac{\tau}{2\tau + 1}\right)^n \tau(3\tau + 1),$$

$$|B_n| = d(A_n, A_{n+2}) = \left(\frac{\tau}{2\tau + 1}\right)^n (3\tau + 1).$$

The intervals  $A_n$  are ordered from left to right  $A_1, A_3, A_5, \dots, A_4, A_2, A_0$ , so any chunk  $P$  of  $S_1$  which does not contain 0 must be a finite union of consecutive even or odd numbered  $A_n$ . Let  $A_n$  be the interval in  $P$  with the largest index; then

$$\frac{|P|}{d(P, S_1 - P)} \geq \frac{|A_n|}{d(A_n, A_{n+2})} = \tau,$$

with equality holding when  $P = A_n$ . On the other hand, if a chunk  $P$  of  $S_1$  contains zero, let  $n$  be the larger index of the leftmost and rightmost  $A_k$  in  $P$ . Then  $P$  must contain  $A_{n-1}$ , and since  $P$  is not all of  $S_1$ ,  $n \geq 2$ , so

$$\frac{|P|}{d(P, S_1 - P)} \geq \frac{|A_n \cup A_{n-1}|}{d(A_n, A_{n-2})} = \frac{(\tau/(2\tau+1))^{n-1}(3\tau+1)(2\tau+1)}{(\tau/(2\tau+1))^{n-2}(3\tau+1)} = \tau.$$

Therefore the thickness of  $S_1$  is  $\tau$ .

Similarly, if  $P$  is a chunk of  $S_2$ , then for an appropriately chosen  $B_n$ , either

$$\frac{|P|}{d(P, S_2 - P)} \geq \frac{|B_n|}{d(B_n, B_{n+2})} = \frac{(2\tau+1)^2}{\tau^3} = g(\tau).$$

or

$$\begin{aligned} \frac{|P|}{d(P, S_2 - P)} &\geq \frac{|B_n \cup B_{n-1}|}{d(B_n, B_{n-2})} \\ &= \frac{(\tau/(2\tau+1))^{n-1}((3\tau+1)/(2\tau+1))(\tau^2+3\tau+1)}{(\tau/(2\tau+1))^{n-2}\tau(3\tau+1)} \\ &= \frac{\tau^2+3\tau+1}{\tau^2} \\ &= f(\tau). \end{aligned}$$

Thus

$$\tau(S_2) = \min(f(\tau), g(\tau)).$$

As we pointed out before, by reducing the thickness of  $S_2$  by an arbitrarily small amount we can shrink the intersection of  $S_1$  and  $S_2$  to a single point. Let  $\tau_1$  denote the thickness of the set  $S_1$ , and let  $\tau_2$  be the thickness of  $S_2$ . Then up to a change of indices, the above construction demonstrates that a single point of intersection can be obtained when either

$$\tau_1 < \min(f(\tau_2), g(\tau_2)) \quad (2.4)$$

or

$$\tau_2 < \min(f(\tau_1), g(\tau_1)). \quad (2.5)$$

Also, if either (2.4) or (2.5) is an equality instead, the intersection can be countable. (Kraft [2] has analyzed this borderline case and determined when the intersection can be finite.) Therefore we can only hope to guarantee an uncountable intersection if

$$\tau_1 > \min(f(\tau_2), g(\tau_2)) \quad (2.6)$$

and

$$\tau_2 > \min(f(\tau_1), g(\tau_1)). \quad (2.7)$$

One may check that  $g(\tau) > f(\tau) > \sqrt{2} + 1$  for  $\tau < \sqrt{2} + 1$  and  $g(\tau) < f(\tau) < \sqrt{2} + 1$  for  $\tau > \sqrt{2} + 1$ . Therefore (2.6) and (2.7) are equivalent to (1.1) in the case  $\tau_1 \geq \tau_2$ , and to (1.2) when  $\tau_2 \geq \tau_1$ .

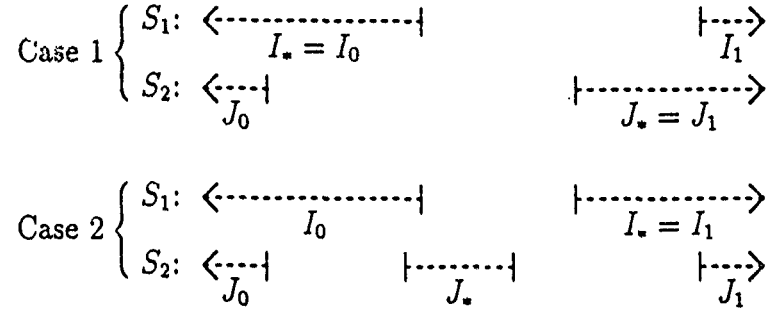


Figure 6: Cases in the construction of  $I_*$  and  $J_*$ .

### 3 Proof of Main Result

We now prove Theorem 1 by constructing a set  $S$  with positive thickness in  $S_1 \cap S_2$ .

**Proof of Theorem 1** Let  $S_1$  and  $S_2$  be interleaved compact sets with  $\tau(S_1) \geq \tau_1$  and  $\tau(S_2) \geq \tau_2$  for some  $(\tau_1, \tau_2)$  in region  $C$  of Figure 3. Let the gaps of  $S_1$  be  $I_0, I_1, I_2, \dots$ , with  $I_0$  and  $I_1$  unbounded,  $I_0$  to the left of  $I_1$ , and  $|I_2| \geq |I_3| \geq \dots$ . For  $S_2$  we define  $J_0, J_1, J_2, \dots$  similarly. We refer to the intervals  $I_n$  and  $J_n$  collectively as the "original gaps". Our goal is to construct the complement of  $S$  as a union of disjoint open intervals  $K_0, K_1, K_2, \dots$  with  $K_0$  and  $K_1$  unbounded, and with every original gap contained in some  $K_m$  (whence  $S \subset S_1 \cap S_2$ .) To get a lower bound on the thickness of  $S$ , observe that every chunk  $P$  of  $S$  is bordered on each side by a gap of  $S$ , with at least one of the bordering gaps being bounded. Pick a chunk  $P$ , and say  $P$  is bordered by  $K_m$  and  $K_n$  with  $m > n$  and  $m \geq 2$ . Then

$$\frac{|P|}{d(P, S - P)} = \frac{d(K_m, K_n)}{\min(|K_m|, |K_n|)} \geq \frac{d(K_m, K_n)}{|K_m|}.$$

The theorem will therefore be proven when we show for some  $\varphi(\tau_1, \tau_2) > 0$  that whenever  $m > n$  and  $m \geq 2$ ,

$$\frac{d(K_m, K_n)}{|K_m|} \geq \varphi(\tau_1, \tau_2). \quad (3.1)$$

We begin by finding a pair of original gaps  $I_*$  and  $J_*$  between which  $S$  will lie; that is,  $I_*$  and  $J_*$  will be contained in  $K_0$  and  $K_1$ . The properties we desire of  $I_*$  and  $J_*$  are that they are a positive distance apart, that all gaps of  $S_1$  with an endpoint between the closures of  $I_*$  and  $J_*$  are bounded and no larger than  $I_*$ , and likewise (in comparison to  $J_*$ ) for gaps of  $S_2$  between  $I_*$  and  $J_*$ . We will show later that once  $I_*$  and  $J_*$  have been determined, the diameter of  $S$  can be bounded below by a constant depending on  $\tau_1$  and  $\tau_2$  times the distance between  $I_*$  and  $J_*$ .

Assume without loss of generality that  $J_0 \subset I_0$ . If  $I_1 \subset J_1$  (Case 1 of Figure 6), then  $I_* = I_0$  and  $J_* = J_1$  have the above properties; they must be separated by a positive distance

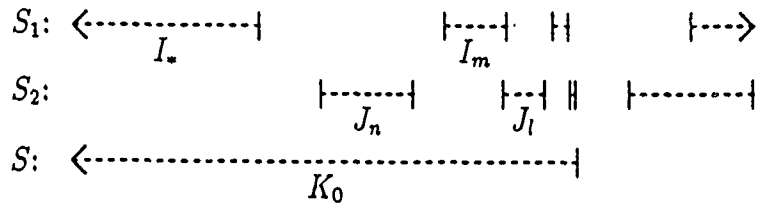


Figure 7: The construction of  $K_0$

since  $S_1$  and  $S_2$  are interleaved. If  $J_1 \subset I_1$  (Case 2 of Figure 6), let  $J_*$  be the largest gap of  $S_2$  with an endpoint between  $I_0$  and  $I_1$ , and let  $I_*$  be whichever of  $I_0$  and  $I_1$  is farthest from  $J_*$ . At least one of  $I_0$  and  $I_1$  must be a positive distance from  $J_*$  since  $S_1$  and  $S_2$  are interleaved.

Next, let  $t$  be a positive constant whose precise value will be chosen later; for now we assume that  $t < (\tau_1\tau_2 - 1)/(\tau_1 + \tau_2 + 2) < \min(\tau_1, \tau_2)$ . Assume without loss of generality that  $I_*$  lies to the left of  $J_*$ . We begin constructing  $K_0$  by requiring that it contain  $I_*$ . We then require that  $K_0$  contain the rightmost bounded  $J_n$  with  $d(I_*, J_n) \leq t|J_n|$  (we will verify that there is a rightmost gap satisfying this condition when we later examine our construction in more detail.) If there does not exist such a  $J_n$  that is not already contained in  $I_*$ , we stop the construction and let  $K_0 = I_*$ . Otherwise, we further require that  $K_0$  contain the rightmost bounded  $I_m$  that is within  $t$  times its length of the previously added  $J_n$ . Again, if this requirement does not extend  $K_0$  any farther rightward, we stop the construction. If not, we then add to  $K_0$  the rightmost  $J_l$  which is within  $t$  times its length of  $I_m$  and is at most as large as  $J_n$  (see Figure 7.) If a next step is necessary, we consider gaps of  $S_1$  which are no larger than  $I_m$ , and so forth. We may have to continue this process infinitely often, but if so we must converge to a right endpoint for  $K_0$ , since there is no way this construction can extend past the rightmost point in  $S_1 \cup S_2$ .

We define  $K_1$  similarly, starting with the requirement that  $K_1$  contain  $J_*$  and extending  $K_1$  to the left if necessary in the same way we constructed  $K_0$ . Next, to construct  $K_2$  we first require that it contain the largest original gap (choose any one in case of a tie) not contained in  $K_0 \cup K_1$  (if no such gap exists, we leave  $K_2$  undefined and let  $S$  be the complement of  $K_0 \cup K_1$ .) Then we extend it on both the left and right in the same manner as before, but considering only gaps that are at most as large as the one we started with, to obtain the endpoints of  $K_2$ . We next start with the largest original gap not contained in  $K_0 \cup K_1 \cup K_2$ , proceeding similarly to define  $K_3$ , and so forth. Any given original gap must eventually be contained in some  $K_n$  because there can be only finitely many original gaps that are as large or larger than the given one. We do not yet know that the  $K_n$  are disjoint from each other; this will follow when we prove (3.1), though.

Let us now examine our construction more closely. Define  $l(I)$  and  $r(I)$  to be respectively

the left and right endpoints of an interval  $I$ . For a given  $K_n$ , let  $G_0$  be the gap we started with in its construction, which for  $n \geq 2$  must be the largest original gap it contains (or at least tied for the largest.) For simplicity we assume here that  $G_0$  is a gap of  $S_1$ . Consider the collection  $E$  of all  $J_n$  with  $|J_n| \leq |G_0|$ ,  $r(J_n) > r(G_0)$ , and  $d(G_0, J_n) \leq t|J_n|$ . We claim that the members of  $E$  (if any) are increasing in size from left to right. If  $J_m, J_n \in E$  with  $J_m$  to the left of  $J_n$ , then since  $r(J_m) < r(J_n)$ , it follows that  $d(J_m, J_n) < d(G_0, J_n) \leq t|J_n|$ . Since  $t < \tau_2$  and  $d(J_m, J_n) \geq \tau_2 \min(|J_m|, |J_n|)$ , it must then be the case that  $|J_n| > |J_m|$ . Thus if  $E$  is not empty, it must have a rightmost member, which we call  $G_1$  (notice that  $G_1$  is also the largest member of  $E$ .) If  $E$  is empty, we let  $G_1$  be empty, but in order to facilitate future formalism, we define  $|G_1| = 0$  and  $r(G_1) = r(G_0)$ . One must keep in mind this degenerate case in verifying the assertions and formulas that follow.

We likewise define  $G_2$  to be the rightmost gap of  $S_1$  which is at most as large as  $G_0$  and lies within  $t$  times its length of  $G_1$ ; again if no such gap exists with  $r(G_2) > r(G_1)$  we say that  $|G_2| = 0$  and  $r(G_2) = r(G_1)$ . Next, to define  $G_3$  we consider only gaps which are at most as large as  $G_1$ , for  $G_4$  we look only at gaps no larger than  $G_2$ , and so forth. Define  $G_{-1}, G_{-2}, \dots$  similarly to be the leftmost (and largest) gaps added to  $K_n$  at each stage of the process of extending  $K_n$  leftward. Then we may think of the open interval  $K_n$  as being defined by

$$l(K_n) = \lim_{m \rightarrow -\infty} l(G_m),$$

$$r(K_n) = \lim_{m \rightarrow \infty} r(G_m).$$

Each limit exists because it is the limit of a bounded monotonic sequence.

In the above construction, the even-numbered  $G_m$  are gaps of  $S_1$  and the odd-numbered ones are gaps of  $S_2$ , but if  $G_0$  had been a gap of  $S_2$  it would be the other way around. In any case,  $G_0$  is the largest even-numbered  $G_m$  and either  $G_1$  or  $G_{-1}$  is the largest odd-numbered one. Also, the even-numbered  $G_m$  decrease monotonically in size as one moves either rightward or leftward from the largest, and the same statement holds for the odd-numbered  $G_m$ . We call a given  $G_m$  either a "1-gap" or "2-gap" of  $K_n$  according to whether it is a gap of  $S_1$  or  $S_2$ . Notice that not all original gaps contained in  $K_n$  are 1-gaps or 2-gaps, only those that have been given a label  $G_m$  in the construction of  $K_n$ . When we refer henceforth to left-to-right ordering or adjacency among the 1-gaps and 2-gaps of a given  $K_n$ , it is with respect to the ordering  $\dots, G_{-2}, G_{-1}, G_0, G_1, G_2, \dots$  (Thus, for instance, 1-gaps can only be adjacent to 2-gaps and vice-versa.)

The following lemma will be used in bounding both the numerator and denominator of the left side of (3.1). It establishes for all  $m \geq 0$  a bound on how far  $K_n$  can extend to the right of  $G_m$  in terms of how far  $G_{m+1}$  extends past  $G_m$ , and similarly for  $m \leq 0$  on the left.

**Lemma 4** Assume  $t < (\tau_1\tau_2 - 1)/(\tau_1 + \tau_2 + 2)$ . Let

$$\sigma_1 = \frac{(\tau_1 - t)(\tau_2 + 1)}{(\tau_1 - t)(\tau_2 - t) - (1 + t)^2}$$

and

$$\sigma_2 = \frac{(\tau_2 - t)(\tau_1 + 1)}{(\tau_1 - t)(\tau_2 - t) - (1 + t)^2}.$$

Let  $G$  be a 1-gap of  $K_n$  which is at least as large as all 1-gaps of  $K_n$  to its right. Let  $H$  be the next 2-gap of  $K_n$  to the right of  $G$ . Then

$$r(K_n) - r(G) \leq \sigma_2(r(H) - r(G)).$$

The same statement with "1" and "2" interchanged holds, as do the corresponding results for left endpoints.

**Proof** Let  $I$  be the next 1-gap of  $K_n$  to the right of  $G$ . Then since  $|I| \leq |G|$ ,

$$\tau_1|I| \leq d(G, I) \leq d(H, I) + r(H) - r(G) \leq t|I| + r(H) - r(G),$$

which, because  $t < \tau_1$ , implies that  $|I|$  is bounded above by  $(r(H) - r(G))/(\tau_1 - t)$ . Hence

$$r(I) - r(H) \leq |I| + d(H, I) \leq (1 + t)|I| \leq \frac{1 + t}{\tau_1 - t}(r(H) - r(G)). \quad (3.2)$$

Likewise the next rightward 2-gap of  $K_n$  extends at most  $((1 + t)/(\tau_2 - t))(r(I) - r(H))$  beyond  $I$ , and by induction

$$\begin{aligned} r(K_n) - r(G) &= r(H) - r(G) + r(I) - r(H) + \cdots \\ &\leq \left(1 + \frac{1 + t}{\tau_1 - t} + \frac{1 + t}{\tau_1 - t} \frac{1 + t}{\tau_2 - t} + \cdots\right) (r(H) - r(G)) \\ &= \sigma_2(r(H) - r(G)). \end{aligned}$$

The geometric series converges, and the denominator of  $\sigma_2$  is positive, because of our assumption that  $t < (\tau_1\tau_2 - 1)/(\tau_1 + \tau_2 + 2)$ . ■

The next lemma builds on Lemma 4 to obtain a positive lower bound on the distance between a given  $K_m$  and  $K_n$ , provided we can find a 1-gap of  $K_m$  and a 2-gap of  $K_n$  which are respectively larger than all 1-gaps and 2-gaps between them. The proof is difficult and will be handled later.

**Lemma 5** There exists a function  $\psi_t(\tau_1, \tau_2)$  that is positive whenever  $(\tau_1, \tau_2)$  is in region  $C$  and  $t$  is sufficiently small, and for which the following statement holds. For  $m \neq n$ , let  $G$  be a 1-gap of  $K_m$  and  $H$  be a 2-gap of  $K_n$ . If all 1-gaps of  $K_m$  or  $K_n$  with at least one endpoint between the closures of  $G$  and  $H$  are bounded and at most as large as  $G$ , and all similarly situated 2-gaps are bounded and at most as large as  $H$ , then

$$d(K_n, K_m) \geq \psi_t(\tau_1, \tau_2)d(G, H).$$

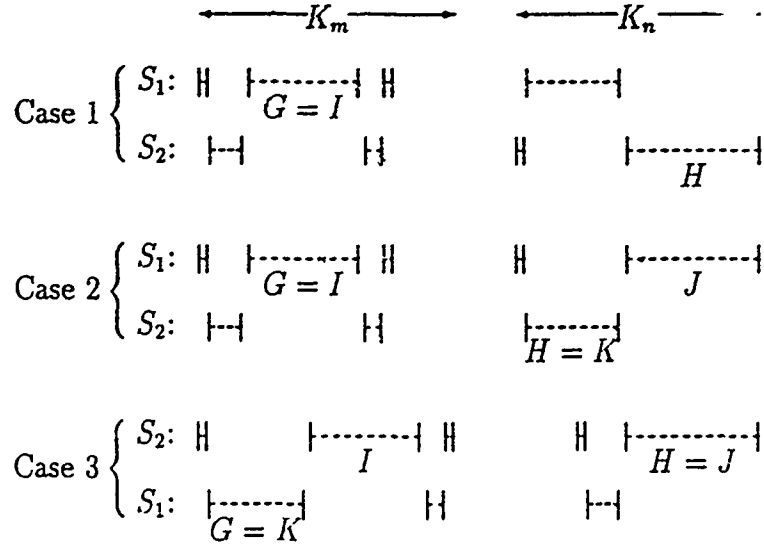


Figure 8: Cases in the proof of (3.1)

Recall that to construct  $K_0$  and  $K_1$ , we chose  $I_*$  and  $J_*$  to satisfy the above hypotheses. Thus we now know that  $K_0$  and  $K_1$  are disjoint and separated by a positive distance (which is at least  $\psi_1(\tau_1, \cdot)$  times the distance between  $I_*$  and  $J_*$ .)

Now suppose  $0 \leq n < m$  and  $m \geq 2$ ; we will prove (3.1) by finding a  $G$  and  $H$  which satisfy the hypotheses of Lemma 5. Assume without loss of generality that  $K_m$  lies to the left of  $K_n$ . Let  $I$  be the largest original gap in  $K_m$ ; say  $I$  is a 1-gap. If all 1-gaps of  $K_n$  are smaller than  $I$  (Case 1 of Figure 8), let  $H$  be the largest original gap in  $K_n$ . Since  $m > n$ ,  $K_n$  was constructed before  $K_m$ , so  $H$  must be at least as large as  $I$ , and thus is a 2-gap. Let  $G = I$ ; then  $G$  and  $H$  satisfy the hypotheses of Lemma 5. Also,  $d(G, H) > t|G|$ , since otherwise  $G$  would have been included in the construction of  $K_n$ . If on the other hand there are 1-gaps of  $K_n$  which are at least as large as  $I$  (Cases 2 and 3 of Figure 8), let  $J$  be the closest such gap to  $I$ . Consider all 2-gaps of  $K_m$  or  $K_n$  to the left of  $J$ ; let  $K$  be the largest such 2-gap (any one will do in case of a tie.) Notice that  $K$  must be adjacent to  $I$  or  $J$ . If  $K$  is in  $K_n$  (Case 2), let  $G = I$  and  $H = K$ ; then  $G$  and  $H$  satisfy the hypotheses of Lemma 5, and  $d(G, H) > t|G|$  because  $G$  was not included in  $K_n$ . Otherwise (Case 3), let  $G = K$  and  $H = J$ , and reverse the indices "1" and "2". Once again,  $G$  and  $H$  satisfy the hypotheses of Lemma 5 and  $d(G, H) > t|G|$ . Notice also that in all cases,  $G$  is the largest 1-gap of  $K_m$ , and  $H$  is at least as large as all 2-gaps of  $K_m$ .

We now estimate how large  $K_m$  can be. Let  $I$  and  $J$  be the 2-gaps of  $K_m$  adjacent to  $G$  on its left and right, respectively. Since  $I$  is at most as large as  $H$ ,

$$\tau_2|I| \leq d(I, H) \leq d(I, G) + |G| + d(G, H) \leq t|I| + |G| + d(G, H),$$

or in other words

$$|I| \leq \frac{1}{\tau_2 - t}(|G| + d(G, H)). \quad (3.3)$$

The same bound holds also for  $J$ , so by Lemma 4,

$$\begin{aligned} |K_m| &= |G| + l(G) - l(K_m) + r(K_m) - r(G) \\ &\leq |G| + \sigma_2(l(G) - l(I)) + \sigma_2(r(J) - r(G)) \\ &\leq |G| + \sigma_2(1 + t)(|I| + |J|) \\ &\leq |G| + 2\sigma_2 \frac{1+t}{(\tau_2 - t)}(|G| + d(G, H)) \\ &\leq \left( \frac{1}{t} + 2\sigma_2 \frac{1+t}{(\tau_2 - t)} \left( \frac{1}{t} + 1 \right) \right) d(G, H) \\ &= \frac{(\tau_1 - t)(\tau_2 - t) + (1 + t)^2(2\tau_1 + 1)}{t((\tau_1 - t)(\tau_2 - t) - (1 + t)^2)} d(G, H). \end{aligned} \quad (3.4)$$

If on the other hand  $G$  is a 2-gap and  $H$  is a 1-gap, we obtain the same bound as (3.4), but with the indices "1" and "2" interchanged. Then in either case,

$$|K_m| \leq \frac{(\tau_1 - t)(\tau_2 - t) + (1 + t)^2(2 \max(\tau_1, \tau_2) + 1)}{t((\tau_1 - t)(\tau_2 - t) - (1 + t)^2)} d(G, H).$$

Finally, by Lemma 5,

$$\frac{d(K_m, K_n)}{|K_m|} \geq \frac{t((\tau_1 - t)(\tau_2 - t) - (1 + t)^2)\psi_t(\tau_1, \tau_2)}{(\tau_1 - t)(\tau_2 - t) + (1 + t)^2(2 \max(\tau_1, \tau_2) + 1)}. \quad (3.5)$$

The right side of (3.5) is positive as long as  $t$  is between 0 and  $(\tau_1\tau_2 - 1)/(\tau_1 + \tau_2 + 2)$ , and  $\psi_t(\tau_1, \tau_2) > 0$ , and goes to zero when  $t$  approaches any of these borderline values. Therefore the right side of (3.5) attains a maximum value, call it  $\varphi(\tau_1, \tau_2)$ , at some allowable value of  $t$ , say  $t_*$ . We thus carry out the construction of  $S$  with  $t = t_*$ ; then (3.1) holds, and the proof is complete. ■

Let  $\psi(\tau_1, \tau_2) = \psi_{t_*}(\tau_1, \tau_2)$ ; then

$$\varphi(\tau_1, \tau_2) = \frac{t_*((\tau_1 - t_*)(\tau_2 - t_*) - (1 + t_*)^2)\psi(\tau_1, \tau_2)}{(\tau_1 - t_*)(\tau_2 - t_*) + (1 + t_*)^2(2 \max(\tau_1, \tau_2) + 1)}.$$

**Remark** We will see in the proof of Lemma 5 that  $\psi(\tau_1, \tau_2)$ , and hence  $\varphi(\tau_1, \tau_2)$ , must be very small when  $(\tau_1, \tau_2)$  is near the boundary of region  $C$ . However, if both  $\tau_1$  and  $\tau_2$  are large and  $t$  is small compared with the two thicknesses, it is not hard to check that  $\psi_t(\tau_1, \tau_2)$  is close to one. Then if  $\tau_1, \tau_2 \gg 1$ , one finds that  $t_*$  is of order  $\sqrt{\min(\tau_1, \tau_2)}$ , whence  $\varphi(\tau_1, \tau_2)$  is of order  $\sqrt{\min(\tau_1, \tau_2)}$  also. Thus when the thicknesses of  $S_1$  and  $S_2$  are large, the lower bound we obtain on the thickness of  $S$  is reasonably large.

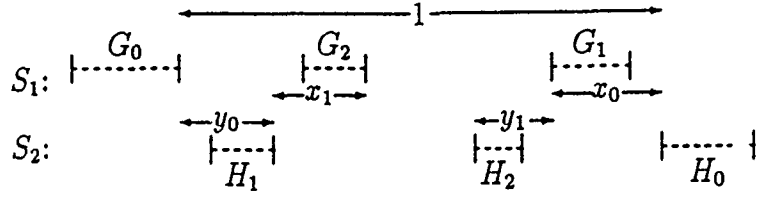


Figure 9: The gaps  $G_i$  and  $H_i$

We now prove our main technical lemma.

**Proof of Lemma 5** Let  $G$  be a 1-gap of  $K_m$  and  $H$  be a 2-gap of  $K_n$  satisfying the hypotheses. We assume without loss of generality that  $\tau_1 \geq \tau_2$ ; then by (1.1) the condition  $(\tau_1, \tau_2) \in C$  implies

$$\tau_1 > f(\tau_2) = \frac{\tau_2^2 + 3\tau_2 + 1}{\tau_2^2} \quad (3.6)$$

and

$$\tau_2 > g(\tau_1) = \frac{(2\tau_1 + 1)^2}{\tau_1^3}. \quad (3.7)$$

If  $d(G, H) = 0$ , the inequality to be proven is trivial. Otherwise, let us normalize  $d(G, H)$  to be one, and assume  $G$  lies to the left of  $H$ . Let  $G_0 = G$  and  $H_0 = H$ . Let  $G_1$  be the 1-gap of  $K_n$  adjacent to  $H_0$  on its left, and let  $H_1$  be the 2-gap of  $K_m$  adjacent to  $G_0$  on its right. Let  $G_2$  be the adjacent 1-gap of  $K_m$  rightward from  $H_1$ , and likewise define  $H_2, G_3, H_3, \dots$  (see Figure 9.) For  $i \geq 0$  let

$$x_i = \begin{cases} l(H_i) - l(G_{i+1}) & i \text{ even} \\ r(G_{i+1}) - r(H_i) & i \text{ odd} \end{cases}$$

and

$$y_i = \begin{cases} r(H_{i+1}) - r(G_i) & i \text{ even} \\ l(G_i) - l(H_{i+1}) & i \text{ odd}. \end{cases}$$

Let  $R_i = d(G_i, H_i)$ ; then  $R_0 = 1$  and  $R_{i+1} = \max(R_i - x_i - y_i, 0)$  for  $i \geq 0$ . Let  $R_\infty$  be the limit as  $i$  goes to infinity of  $R_i$ . Then  $d(K_m, K_n) = R_\infty$ , so we wish to show that there is a positive lower bound on  $R_\infty$  which depends only on  $\tau_1, \tau_2$ , and  $t$ .

In the same way as we obtained (3.2) it follows that for all  $i$ ,

$$x_{i+1} \leq \frac{1+t}{\tau_1-t} y_i \quad (3.8)$$

and

$$y_{i+1} \leq \frac{1+t}{\tau_2-t} x_i. \quad (3.9)$$

Furthermore, by Lemma 4 we have that

$$y_i + x_{i+1} + y_{i+2} + \dots \leq \sigma_2 y_i$$

and

$$x_i + y_{i+1} + x_{i+2} + \cdots \leq \sigma_1 x_i.$$

Thus, for each  $i$ ,

$$R_\infty \geq R_i - x_i - y_i - x_{i+1} - y_{i+1} - \cdots \geq R_i - \sigma_1 x_i - \sigma_2 y_i. \quad (3.10)$$

We will show that for some  $i$ , the right side of (3.10) is positive.

Next let us obtain upper bounds on  $x_0$  and  $y_0$ . We know that

$$x_0 = l(H_0) - l(G_1) \leq |G_1| + d(G_1, H_0) \leq (1+t)|G_1|, \quad (3.11)$$

and by hypothesis  $|G_1| \leq |G_0|$ , so

$$x_0 = l(H_0) - r(G_0) - (l(G_1) - r(G_0)) = 1 - d(G_0, G_1) \leq 1 - \tau_1 |G_1|. \quad (3.12)$$

Eliminating  $|G_1|$  from these inequalities yields

$$x_0 \leq \frac{1+t}{\tau_1 + 1 + t}. \quad (3.13)$$

Similarly,

$$y_0 \leq \frac{1+t}{\tau_2 + 1 + t}. \quad (3.14)$$

We can obtain similar bounds on  $x_i$  and  $y_i$  for  $i \geq 1$ , but the bounds are complicated by the fact that we do not know in general that  $|G_{i+1}| \leq |G_i|$  (or  $|H_{i+1}| \leq |H_i|$ ). The analogues of (3.11) and (3.12) are thus

$$x_i \leq (1+t)|G_{i+1}|$$

and

$$x_i \leq R_i - \tau_1 \min(|G_i|, |G_{i+1}|). \quad (3.15)$$

If  $|G_{i+1}| \leq |G_i|$ , then as in (3.13) it follows that

$$x_i \leq \frac{1+t}{\tau_1 + 1 + t} R_i. \quad (3.16)$$

If  $|G_{i+1}| > |G_i|$ , then by (3.15),

$$x_i \leq R_i - \tau_1 |G_i| \leq R_i - \frac{\tau_1}{1+t} x_{i-1}. \quad (3.17)$$

If (3.16) fails, then using (3.17) together with the negation of (3.16), one finds that  $x_{i-1}$  is bounded above by the right side of (3.16). Thus regardless of the relative lengths of  $G_i$  and  $G_{i+1}$ ,

$$\min(x_i, x_{i-1}) \leq \frac{1+t}{\tau_1 + 1 + t} R_i. \quad (3.18)$$

for  $i \geq 1$ . Likewise, regardless of the relative lengths of  $H_i$  and  $H_{i+1}$ , we have for all  $i \geq 1$  that

$$\min(y_i, y_{i-1}) \leq \frac{1+t}{\tau_2+1+t} R_i. \quad (3.19)$$

Let  $a_i = x_i/R_i$  and  $b_i = y_i/R_i$  provided  $R_i > 0$ ; then

$$R_{i+1} = \max(1 - a_i - b_i, 0) R_i.$$

Thus  $a_{i+1}$  and  $b_{i+1}$  are defined as long as  $1 - a_i - b_i > 0$ . For  $j = 1, 2$  let

$$\lambda_j = \frac{1+t}{\tau_j+1+t}, \quad (3.20)$$

$$\mu_j = \frac{1+t}{\tau_j-t}.$$

The conditions (3.13), (3.14), (3.18), and (3.19) can then be written

$$a_0 \leq \lambda_1,$$

$$b_0 \leq \lambda_2,$$

$$\min\left(a_{i+1}, \frac{a_i}{1-a_i-b_i}\right) \leq \lambda_1, \quad (3.21)$$

and

$$\min\left(b_{i+1}, \frac{b_i}{1-a_i-b_i}\right) \leq \lambda_2.$$

Also, conditions (3.8) and (3.9) become

$$a_{i+1} \leq \mu_1 \frac{b_i}{1-a_i-b_i}$$

and

$$b_{i+1} \leq \mu_2 \frac{a_i}{1-a_i-b_i}. \quad (3.22)$$

Finally, our objective is to show that for some  $i$ ,

$$1 - \sigma_1 a_i - \sigma_2 b_i > 0, \quad (3.23)$$

which implies that the right side of (3.10) is positive.

We observe that  $a_{i+1}$  and  $b_{i+1}$  are defined at least as long as  $a_i \leq \lambda_1$  and  $b_i \leq \lambda_2$ , because then

$$\begin{aligned} 1 - a_i - b_i &\geq 1 - \lambda_1 - \lambda_2 \\ &= \frac{\tau_1 \tau_2 - (1+t)^2}{(\tau_1 + t + 1)(\tau_2 + t + 1)} \\ &> \frac{(\tau_1 - t)(\tau_2 - t) - (1+t)^2}{(\tau_1 + t + 1)(\tau_2 + t + 1)} \\ &> 0 \end{aligned}$$

(since  $t < (\tau_1 \tau_2 - 1)/(\tau_1 + \tau_2 + 2)$ .) Also, as long as  $a_i \leq \lambda_1$ , by (3.22) we have

$$b_{i+1} \leq \frac{\mu_2 \lambda_1}{1 - \lambda_1 - b_i}.$$

Let

$$h(b) = \frac{\mu_2 \lambda_1}{1 - \lambda_1 - b}.$$

The equation  $h(b) = b$  has two solutions,

$$b_{\pm} = \frac{1 - \lambda_1 \pm \sqrt{(1 - \lambda_1)^2 - 4\mu_2 \lambda_1}}{2},$$

and if the roots are real, then  $h(b) < b$  for  $b_- < b < b_+$  (this can be verified by checking the value  $b = (1 - \lambda_1)/2$ .) We claim that for  $t$  sufficiently small,  $b_{\pm}$  are real, with

$$b_+ > \lambda_2 \quad (3.24)$$

and

$$1 - \sigma_1 \lambda_1 - \sigma_2 b_- > 0. \quad (3.25)$$

Let us delay the verification of this claim until the end of the proof. Choose  $b_* > b_-$  with  $1 - \sigma_1 \lambda_1 - \sigma_2 b_* > 0$ . Now  $b_0 \leq \lambda_2 < b_+$ , and as long as  $a_i \leq \lambda_1$  continues to hold,  $b_{i+1} \leq h(b_i) < b_i$  for  $b_i \in (b_-, b_+)$ . Then eventually  $b_i \leq b_*$ , and furthermore since  $b - h(b)$  must have a positive minimum value on  $[b_*, \lambda_2]$  (if  $b_* > \lambda_2$  then  $b_0 < b_*$  already) there is a maximum number  $N$  (depending only on  $\tau_1, \tau_2$ , and  $t$ ) of iterations it can take before  $b_i \leq b_*$ . We therefore have shown that if  $a_i \leq \lambda_1$  for  $i \leq N$ , then  $b_i \leq b_*$  for some  $i \leq N$ , and hence

$$1 - \sigma_1 a_i - \sigma_2 b_i \geq 1 - \sigma_1 \lambda_1 - \sigma_2 b_* > 0. \quad (3.26)$$

If on the other hand  $a_{i+1} > \lambda_1$  for some  $i \leq N$ , then let  $i$  be the smallest index for which this occurs. We claim that then (3.23) holds for  $i$ . By the results of the previous paragraph,  $b_i < b_{i-1} < \dots < b_0 \leq \lambda_2$ . Also, by (3.21),  $a_i \leq \lambda_1(1 - a_i - b_i)$ , or in other words

$$a_i \leq \frac{\lambda_1}{1 + \lambda_1}(1 - b_i).$$

Then

$$1 - \sigma_1 a_i - \sigma_2 b_i \geq 1 - \frac{\sigma_1 \lambda_1}{1 + \lambda_1} - \left( \sigma_2 - \frac{\sigma_1 \lambda_1}{1 + \lambda_1} \right) b_i.$$

Now when  $t = 0$ ,

$$\begin{aligned} \sigma_2 - \frac{\sigma_1 \lambda_1}{1 + \lambda_1} &= \frac{(\tau_1 + 1)\tau_2}{\tau_1 \tau_2 - 1} - \frac{\tau_1(\tau_2 + 1)}{(\tau_1 + 2)(\tau_1 \tau_2 - 1)} \\ &= \frac{\tau_1(\tau_1 \tau_2 - 1) + 2\tau_1 \tau_2 + 2\tau_2}{(\tau_1 + 2)(\tau_1 \tau_2 - 1)} \\ &> 0, \end{aligned}$$

and thus for  $t$  sufficiently small it remains positive. Then since  $b_i \leq \lambda_2$ ,

$$\begin{aligned} 1 - \sigma_1 a_i - \sigma_2 b_i &\geq 1 - \frac{\sigma_1 \lambda_1}{1 + \lambda_1} - \left( \sigma_2 - \frac{\sigma_1 \lambda_1}{1 + \lambda_1} \right) \lambda_2 \\ &= 1 - \sigma_2 \lambda_2 - (1 - \lambda_2) \frac{\sigma_1 \lambda_1}{1 + \lambda_1}. \end{aligned} \quad (3.27)$$

When  $t = 0$ , by (3.6)

$$\begin{aligned} 1 - \sigma_2 \lambda_2 &= 1 - \frac{(\tau_1 + 1)\tau_2}{(\tau_2 + 1)(\tau_1 \tau_2 - 1)} \\ &= \frac{\tau_1 \tau_2^2 - 2\tau_2 - 1}{(\tau_2 + 1)(\tau_1 \tau_2 - 1)} \\ &> \frac{\tau_2^2 + 3\tau_2 + 1 - 2\tau_2 - 1}{(\tau_2 + 1)(\tau_1 \tau_2 - 1)} \\ &= \frac{\tau_2}{\tau_1 \tau_2 - 1}, \end{aligned}$$

while

$$(1 - \lambda_2) \frac{\sigma_1 \lambda_1}{1 + \lambda_1} = \frac{\tau_2}{\tau_2 + 1} \cdot \frac{\tau_1(\tau_2 + 1)}{(\tau_1 + 2)(\tau_1 \tau_2 - 1)} = \frac{\tau_1}{\tau_1 + 2} \cdot \frac{\tau_2}{\tau_1 \tau_2 - 1},$$

so the right side of (3.27) is positive for  $t = 0$ . It therefore remains positive for  $t$  sufficiently small.

To summarize, we have shown that if  $t$  is sufficiently small, then for some  $i \leq N$ , either (3.26) or (3.27) holds. The right side of each of these equations is positive and depends only on  $\tau_1, \tau_2$ , and  $t$ . Furthermore,  $a_j \leq \lambda_1$  and  $b_j \leq \lambda_2$  for  $j \leq i$ , so by (3.20),  $R_{j+1} \geq (1 - \lambda_1 - \lambda_2)R_j$ , and hence  $R_i \geq (1 - \lambda_1 - \lambda_2)^i$ . Then by (3.10),

$$R_\infty \geq R_i(1 - \sigma_1 a_i - \sigma_2 b_i) \geq (1 - \lambda_1 - \lambda_2)^N(1 - \sigma_1 a_i - \sigma_2 b_i),$$

where  $1 - \sigma_1 a_i - \sigma_2 b_i$  is in turn bounded below by the lesser of the right sides of (3.26) and (3.27). We have therefore shown for  $t$  sufficiently small how to obtain a positive lower bound on  $R_\infty$  which depends only on  $\tau_1, \tau_2$ , and  $t$ ; we let  $\psi_t(\tau_1, \tau_2)$  be this lower bound.

It remains for us to verify (3.24) and (3.25). We again show they are true for  $t = 0$ , whence they hold for  $t$  sufficiently small by continuity. When  $t = 0$ ,

$$\begin{aligned} b_\pm &= \frac{\tau_1/(\tau_1 + 1) \pm \sqrt{\tau_1^2/(\tau_1 + 1)^2 - 4/((\tau_1 + 1)\tau_2)}}{2} \\ &= \frac{\tau_1 \tau_2 \pm \sqrt{\tau_1^2 \tau_2^2 - 4(\tau_1 + 1)\tau_2}}{2(\tau_1 + 1)\tau_2}. \end{aligned} \quad (3.28)$$

Now by (3.6),

$$\tau_1^2 \tau_2^2 - 4(\tau_1 + 1)\tau_2 = \tau_1(\tau_1 \tau_2^2 - 4\tau_2) - 4\tau_2$$

$$\begin{aligned}
&> \tau_1(\tau_2^2 - \tau_2 + 1) - 4\tau_2 \\
&> \frac{(\tau_2^2 + 3\tau_2 + 1)(\tau_2^2 - \tau_2 + 1) - 4\tau_2^3}{\tau_2^2} \\
&= \frac{(\tau_2^2 - \tau_2 - 1)^2}{\tau_2^2}.
\end{aligned} \tag{3.29}$$

Thus  $b_{\pm}$  are real and distinct (and the same must then hold for  $t$  sufficiently small.) Next by (3.28),

$$b_+ = \frac{\tau_2 + \sqrt{\tau_2^2 - 4\tau_2(1/\tau_1 + 1/\tau_1^2)}}{2\tau_2(1 + 1/\tau_1)}$$

from which we see that  $b_+$  is increasing as a function of  $\tau_1$ . Thus  $b_+$  is greater than the value it would take on if (3.6) were an equality, which owing to (3.29) means

$$\begin{aligned}
b_+ &> \frac{(\tau_2^2 + 3\tau_2 + 1)/\tau_2 + |\tau_2^2 - \tau_2 - 1|/\tau_2}{2(2\tau_2^2 + 3\tau_2 + 1)/\tau_2} \\
&\geq \frac{\tau_2^2 + 3\tau_2 + 1 - (\tau_2^2 - \tau_2 - 1)}{2(\tau_2 + 1)(2\tau_2 + 1)} \\
&= \frac{1}{\tau_2 + 1} \\
&= \lambda_2.
\end{aligned}$$

Hence (3.24) holds for  $t = 0$ , and consequently for  $t$  sufficiently small.

When  $t = 0$ , (3.25) can be written

$$b_- < \frac{1 - \sigma_1 \lambda_1}{\sigma_2} = \frac{\tau_1^2 \tau_2 - 2\tau_1 - 1}{(\tau_1 + 1)^2 \tau_2}. \tag{3.30}$$

The right side of (3.30) is an increasing function of  $\tau_2$ , and since

$$b_- = \frac{\tau_1 - \sqrt{\tau_1^2 - 4(\tau_1 + 1)/\tau_2}}{2(\tau_1 + 1)},$$

$b_-$  is a decreasing function of  $\tau_2$ . Then by (3.7),

$$\begin{aligned}
b_- &< \frac{\tau_1 - \sqrt{\tau_1^2 - 4\tau_1^3(\tau_1 + 1)/(2\tau_1 + 1)^2}}{2(\tau_1 + 1)} \\
&= \frac{\tau_1((2\tau_1 + 1) - \sqrt{(2\tau_1 + 1)^2 - 4(\tau_1^2 + \tau_1)})}{2(\tau_1 + 1)(2\tau_1 + 1)} \\
&= \frac{\tau_1^2}{(\tau_1 + 1)(2\tau_1 + 1)},
\end{aligned}$$

while

$$\frac{\tau_1^2 \tau_2 - 2\tau_1 - 1}{(\tau_1 + 1)^2 \tau_2} > \frac{(2\tau_1 + 1)^2/\tau_1 - (2\tau_1 + 1)}{(\tau_1 + 1)^2(2\tau_1 + 1)^2/\tau_1^3}$$

$$\begin{aligned}
&= \frac{\tau_1^2(2\tau_1 + 1 - \tau_1)}{(\tau_1 + 1)^2(2\tau_1 + 1)} \\
&= \frac{\tau_1^2}{(\tau_1 + 1)(2\tau_1 + 1)}.
\end{aligned}$$

Thus (3.25) holds for  $t = 0$ , and for  $t$  sufficiently small. The proof of Lemma 5 is now complete. ■

## 4 Intersecting Three or More Cantor Sets

In proving Theorem 1, we chose a subset  $S$  of  $S_1 \cap S_2$  in order to guarantee positive thickness. In this section we demonstrate that positive thickness sets are in some sense generic in  $S_1 \cap S_2$ . We also explain how Theorem 1 is useful in finding conditions under which three or more Cantor sets must have a nonempty intersection.

The set  $S$  we constructed in Section 3 need not be dense in  $S_1 \cap S_2$  nor even in the non-isolated points of  $S_1 \cap S_2$ . However, there are subsets of  $S_1 \cap S_2$  with thickness at least  $\varphi(\tau_1, \tau_2)$  near any accumulation point. To see this, let  $\{q_n\}$  be a sequence of distinct points in  $S_1 \cap S_2$  which converge to a point  $q$ . It is not hard to show that within any neighborhood  $N$  of  $q$  there are compact subsets  $T_1 \subset S_1$  and  $T_2 \subset S_2$ , each of which contains all but finitely many  $q_n$ , with  $\tau(T_1) \geq \tau(S_1)$  and  $\tau(T_2) \geq \tau(S_2)$ . Notice that any two compact sets which intersect in three or more points must be interleaved. Thus  $T_1$  and  $T_2$  are interleaved, and by Theorem 1 their intersection contains a set with thickness at least  $\varphi(\tau_1, \tau_2)$ . We conclude that arbitrarily near any non-isolated point of  $S_1 \cap S_2$  there are subsets of  $S_1 \cap S_2$  which have thickness at least  $\varphi(\tau_1, \tau_2)$ .

In addition to showing that there are many subsets of  $S_1 \cap S_2$  with positive thickness, it is possible to obtain a lower bound on the diameter of the positive thickness subset  $S$  of  $S_1 \cap S_2$ . If the two sets  $S_1$  and  $S_2$  are interleaved in such a way that neither is contained in the convex hull of the other, then by the discussion following the statement of Lemma 5, the diameter of  $S$  is at least  $\psi(\tau_1, \tau_2)$  times the length of overlap between the convex hulls of  $S_1$  and  $S_2$ . Since the thickness of  $S$  is at least  $\varphi(\tau_1, \tau_2)$ , we immediately have the following result.

**Corollary 6** *Let  $S_1$  and  $S_2$  be two interleaved compact sets whose thicknesses  $(\tau_1, \tau_2)$  lie in region  $C$  and for which the intersection  $Q$  of their convex hulls contains neither  $S_1$  nor  $S_2$ . If  $S_3$  is a compact set with largest bounded gap  $G$  such that*

- (i) *the hull of  $S_3$  contains  $Q$ ,*
- (ii)  $|G| < \psi(\tau_1, \tau_2)|Q|$
- (iii)  $\tau(S_3)\varphi(\tau_1, \tau_2) \geq 1,$

then  $S_1 \cap S_2 \cap S_3$  is nonempty.

We note that if instead of condition (iii) we required the pair  $\tau(S_3)$  and  $\varphi(\tau_1, \tau_2)$  to lie in  $C$ , then  $S_1 \cap S_2 \cap S_3$  would contain a set of thickness at least  $\varphi(\tau(S_3), \varphi(\tau_1, \tau_2))$ . Thus one can inductively find thickness conditions guaranteeing the nonempty intersection of any finite (or even countably infinite) number of compact sets, although the analogue of the interleaving condition gets more complicated.

If  $(\tau_1, \tau_2)$  is sufficiently far from the boundary of region  $C$ , then as discussed in the remark preceding the proof of Lemma 5 it is not hard to obtain explicit lower bounds on  $\varphi(\tau_1, \tau_2)$  and  $\psi(\tau_1, \tau_2)$ . In particular, for  $\tau_1$  and  $\tau_2$  large we found that  $\varphi(\tau_1, \tau_2)$  is at least of order  $\sqrt{\min(\tau_1, \tau_2)}$ , and  $\psi(\tau_1, \tau_2)$  is approximately one.

We thank the referee for a thorough reading of our paper and many helpful comments.

## References

- [1] K. J. Falconer, The Geometry of Fractal Sets, (Cambridge University Press, 1985).
- [2] R. Kraft, "Intersections of thick Cantor sets," Mem. Amer. Math. Soc. (to appear).
- [3] J. M. Marstrand, "Some fundamental geometrical properties of plane sets of fractional dimensions," Proc. London Math. Soc. (3) 4 (1954), 257-302.
- [4] P. Mattila, "Hausdorff dimension and capacities of intersections of sets in  $n$ -space," Acta Math. 152 (1984), 77-105.
- [5] S. Newhouse, "Nondensity of axiom  $A(a)$  on  $S^2$ ," Proc. A.M.S. Symp. in Pure Math. 14 (1970), 191-202.
- [6] S. Newhouse, "Diffeomorphisms with infinitely many sinks," Topology 13 (1974), 9-18.
- [7] S. Newhouse, "The abundance of wild hyperbolic sets and non-smooth stable sets for diffeomorphisms," Publ. Math. IHES 50 (1979), 101-151.
- [8] S. Newhouse, "Lectures on Dynamical Systems," Progress in Math. 8 (Boston: Birkhauser, 1980), 1-114.
- [9] C. Robinson, "Bifurcation to infinitely many sinks," Comm. Math. Phys. 90 (1983), 433-459.
- [10] R. F. Williams, "How big is the intersection of two thick Cantor sets?", to appear in M. Brown, ed., Contemporary Mathematics, Proceedings of the 1989 Joint Summer Research Conference on Continua and Dynamics.

BORDER-COLLISION BIFURCATIONS  
INCLUDING 'PERIOD TWO TO PERIOD THREE'  
FOR PIECEWISE SMOOTH SYSTEMS\*

by

Helena E. Nusse<sup>a,b</sup>

James A. Yorke<sup>a,c</sup>

December 1990

\* Research in part supported by the Department of Energy (Scientific Computing Staff Office of Energy Research), and by DARPA/ONR.

a. University of Maryland, Institute for Physical Science and Technology, College Park, MD 20742, U.S.A.

b. Rijksuniversiteit Groningen, F.E.W., Vakgroep Econometrie, Postbus 800, NL-9700 AV Groningen, The Netherlands

c. University of Maryland, Department of Mathematics, College Park, MD 20742, U.S.A.

ABSTRACT. We examine bifurcation phenomena for maps that are piecewise smooth and depend continuously on a parameter  $\mu$ . In the simplest case there is a surface  $\Gamma$  in phase space along which the map has no derivative (or has two one-sided derivatives).  $\Gamma$  is the border of two regions in which the map is smooth. As the parameter  $\mu$  is varied, a fixed point  $E_\mu$  may collide with the border  $\Gamma$ , and we may assume that this collision occurs at  $\mu = 0$ . A variety of bifurcations occur frequently in such situations, but never or almost never occur in smooth systems. In particular  $E_\mu$  may cross the border and so will exist for  $\mu < 0$  and for  $\mu > 0$  but may be a saddle in one case, say  $\mu < 0$ , and may be a repeller for  $\mu > 0$ . For  $\mu < 0$  there can be a stable period two orbit which shrinks to the point  $E_0$  as  $\mu \rightarrow 0$ , and for  $\mu > 0$  there may be a stable period 3 orbit which similarly shrinks to  $E_0$  as  $\mu \rightarrow 0$ . Hence one observes the following stable periodic orbits: a stable period 2 orbit collapses to a point and is reborn as a stable period 3 orbit. We also see analogously "stable period 2 to stable period  $p$  orbit bifurcations", with  $p = 5, 11, 52$ , or period 2 to quasi-periodic or even to a chaotic attractor. We believe this phenomenon will be seen in many applications.

## 1. INTRODUCTION

Certain bifurcation phenomena have been reported repeatedly in numerous studies of low dimensional dynamical systems, that depend on one parameter. The rather familiar bifurcation phenomena describing the evolution of attractors as a parameter is varied include the saddle node bifurcation, the period doubling (or halving) bifurcation, and the Hopf bifurcation. In the literature

dealing with bifurcation theory, it is frequently assumed that the map corresponding to the dynamical system is differentiable; see for example [GH], [K], [R], or [S]. To remind the reader so that we may draw contrasts, the well known bifurcation diagram of the quadratic map  $Q_\mu(x) = \mu - x^2$  is given in Figure 1 ( $1 < \mu < 1.5$ ). All the computer assisted pictures were made by using the DYNAMICS program [Y].

FIGURE 1

We say a map is smooth if the map has a continuous derivative. A region is a closed, connected subset in phase space. We examine continuous maps which are piecewise smooth. We restrict attention to those which are smooth on two regions of the plane with the border between these regions being a smooth curve. From now on we assume that there is a smooth curve  $\Gamma$  which separates the plane into two regions denoted by  $R_A$  and  $R_B$ . We say, a map  $F$  from the phase space  $\mathbb{R}^2$  to itself is piecewise-smooth if (1)  $F$  is continuous, and (2)  $F$  is smooth on both the regions  $R_A$  and  $R_B$ . Note that on the border  $\Gamma$  between the regions, the mappings must be equal since  $F_\mu$  is assumed to be continuous. A special case that we shall refer to frequently is the following prototype example, a piecewise linear map into which other generic piecewise linear maps in the plane can be transformed by changes in coordinates.

Let  $u$  and  $w$  be vectors in the plane. Let  $x$  and  $y$  be the phase space coordinates and  $\mu$  is a scalar parameter. Let  $P_\mu$  be the map defined by

$$P_\mu(x,y) = xu + |x|w + (y + \mu)(1,0)$$

and we investigate trajectories  $(x_{n+1}, y_{n+1}) = P_\mu(x_n, y_n)$ . The regions  $R_A$  and  $R_B$  are the left and right half plane separated by

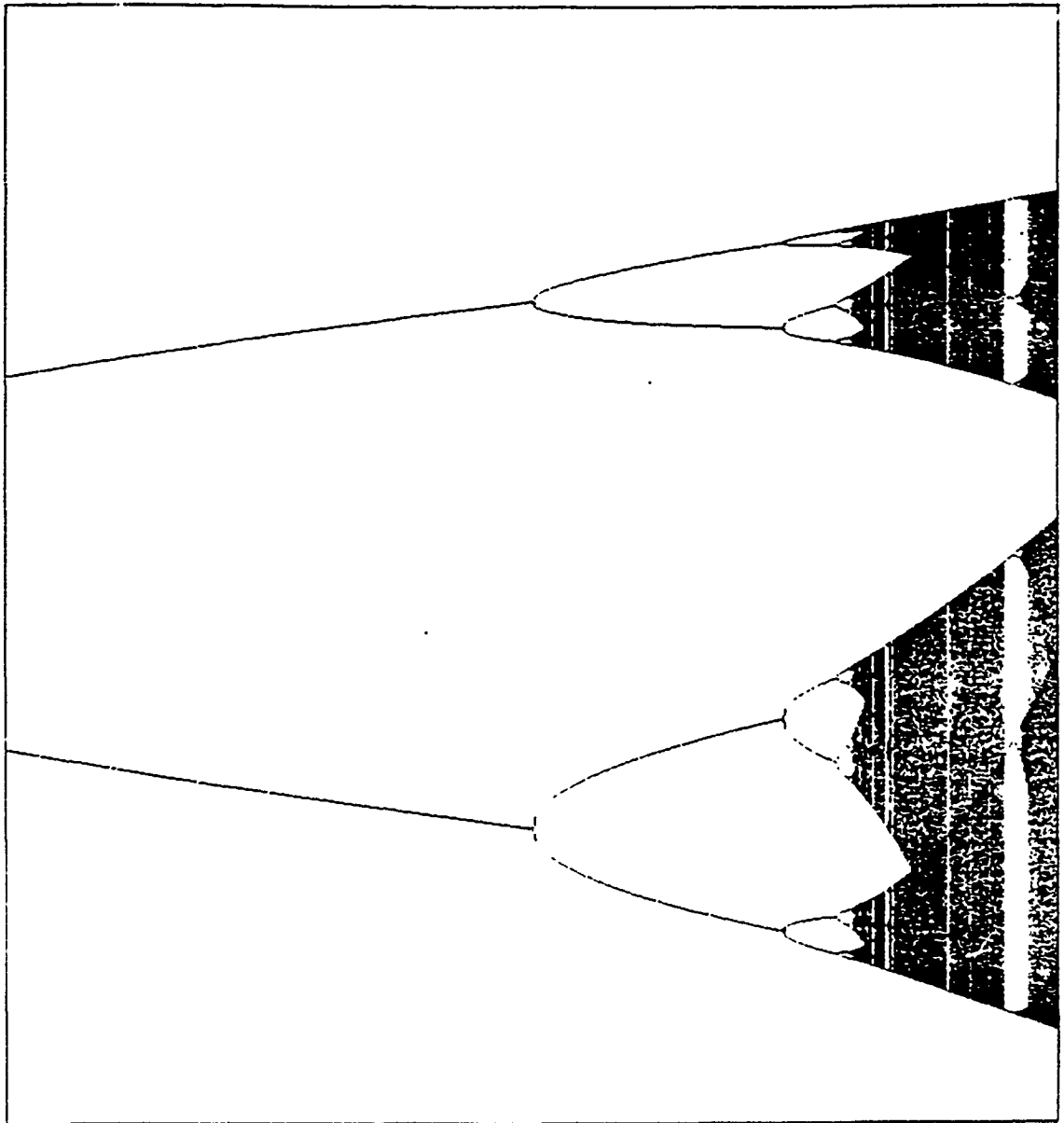


Figure 1.

Bifurcation diagram of the quadratic map  $Q_\mu(x) = \mu - x^2$ . The parameter  $\mu$  (plotted horizontally) varies from 1 to 1.5, and  $x$  is plotted vertically,  $-1 \leq x \leq 2$ .

$\Gamma$ , the Y-axis.

To illustrate the "period two to period three" border-collision bifurcation phenomenon, consider the one-parameter family of maps  $f_\mu$  ( $-\infty < \mu < \infty$ ) from the plane to itself, defined by

$$f_\mu(x,y) = \begin{cases} (-1.4x + y, -0.1x) + \mu(1,0) & \text{if } x \leq 0 \\ (-3x + y, -4x) + \mu(1,0) & \text{if } x \geq 0 \end{cases}$$

Notice that the map is smooth in each of the half planes  $x \leq 0$  and  $x \geq 0$ , and the Y-axis is the border which is a smooth curve. Note that to write  $f_\mu$  in the form of  $P_\mu$ , let  $u = (-2.2, -2.05)$ , and  $w = (-0.8, -1.95)$ . The bifurcation diagram exhibiting the "period two to period three" bifurcation, is presented in Figure 2 ( $-0.1 < \mu < 0.2$ ). All the bifurcation diagrams in this paper show a projection of the attractor, projecting  $(x,y)$  onto the X-axis, which is plotted vertically; the horizontal coordinate is  $\mu$ .

FIGURE 2

The purpose of this paper is to study the occurrence of such a new bifurcation phenomenon for continuous, piecewise smooth maps. These systems include, for example, two-dimensional continuous, piecewise-linear maps. In [HNS] the dynamics of a simple economic model was studied, and a "period three to period two" bifurcation was observed numerically, and was established rigorously in [HN] for a degenerate piecewise-linear situation. The "border-collision bifurcation" phenomena is a much richer class of bifurcation phenomena than just a "period two to period three" bifurcation and occur for generic piecewise smooth maps. We present phenomena that occurs when the nature of an unstable fixed point of a piecewise smooth map is changed while the fixed point collides with the

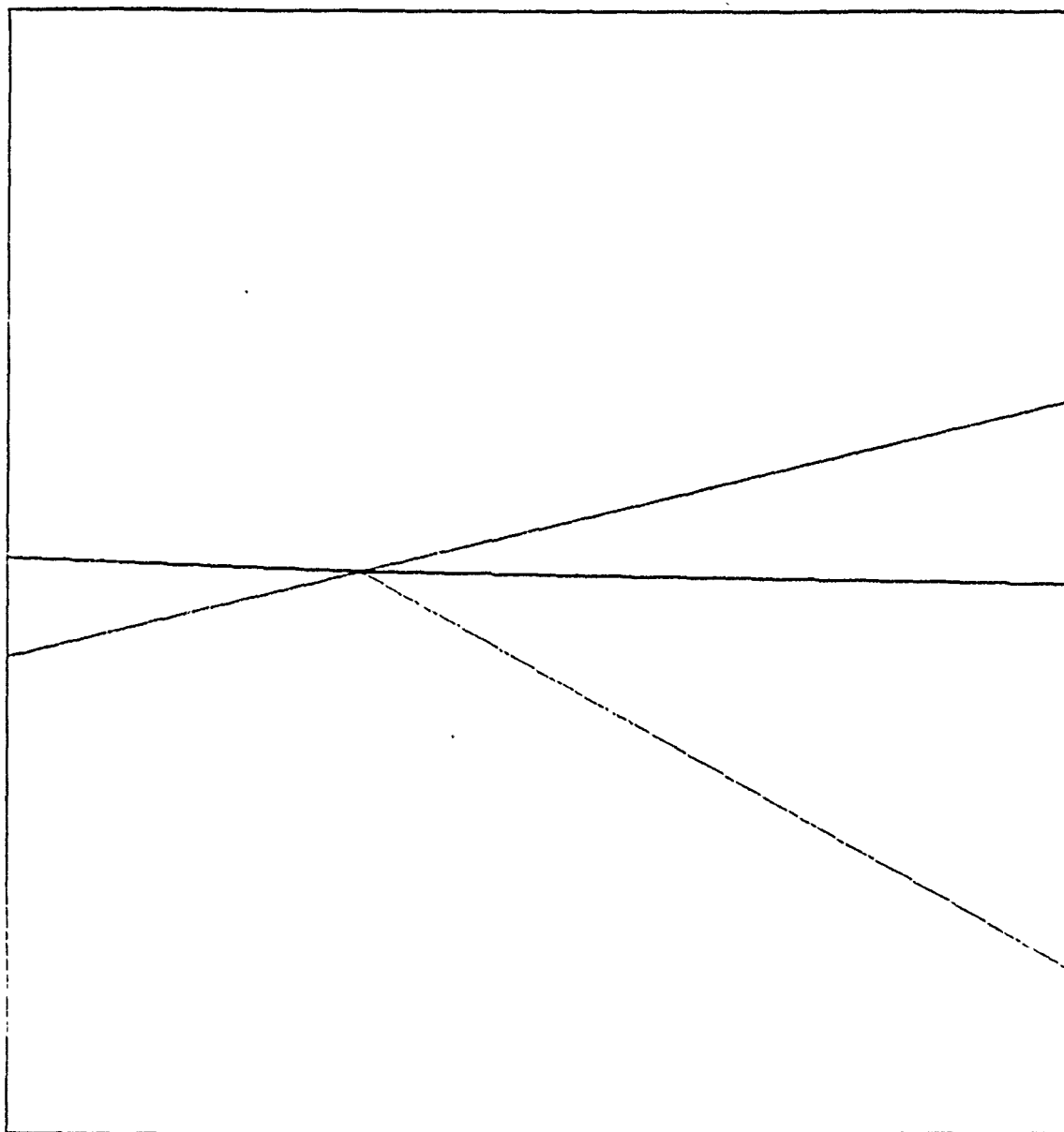


Figure 2.

Bifurcation diagram exhibiting the "period two to period three" bifurcation of the map

$$f_{\mu}(x,y) = (-1.4x + y + \mu, -0.1x) \text{ if } x \leq 0, \text{ and} \\ = (-3x + y + \mu, -4x) \text{ if } x > 0.$$

The parameter  $\mu$  (plotted horizontally) varies from -0.1 to 0.2, and the coordinate  $x$  is plotted vertically,  $-1 \leq x \leq 1$ .

border between two regions in which the map is smooth. Since the fixed point is unstable before and after collision, it is not shown in the bifurcation diagram in Figure 2. While we consider maps in the plane, higher dimensional analogues exist. We know of no phenomena that can occur only in higher dimensional cases. There is also no difficulty in changing the notation to that there are more than 2 regions on which the map is smooth. We could also allow  $\Gamma$  to depend on  $\mu$ , but coordinates could be chosen so that it remains fixed, so our case in practice includes moving boundaries. With moving boundaries the map would be piecewise smooth in  $\mu$ .

We say, a fixed point  $E_\mu$  is a border crossing fixed point if it crosses the border  $\Gamma$  between two regions in which the map is smooth. We will assume that the crossing occurs at  $\mu = 0$ . The fixed point  $E_\mu$  is called a flip saddle if the eigenvalues  $\lambda$  and  $\nu$  of the Jacobian matrix  $DF_\mu(E_\mu)$  if  $\lambda < -1 < \nu < 1$ . Assume that there exists a one-parameter family of piecewise smooth maps and assume that there is a border crossing fixed point (or periodic point)  $E_\mu$ , we emphasize the case when  $E_\mu$  crosses the border  $\Gamma$  it changes from being a flip saddle to a repellor with complex eigenvalues. The above example has this behavior.

In Section 2, we discuss why the border-collision bifurcation phenomenon occurs when the nature of an unstable equilibrium changes when it crosses the border of two regions. To be somewhat more specific, assume that a border crossing fixed point (or periodic point)  $E_\mu$  of a one-parameter family of piecewise smooth maps changes from being a flip saddle to a repellor with complex eigenvalues when it crosses the border  $\Gamma$ . Then at  $\mu = 0$ , a border-collision-bifurcation occurs at this fixed point  $E_\mu$  on the border.

In Section 3, we mainly deal with two piecewise smooth systems of the plane, one piecewise linear and one piecewise nonlinear. The first system is the map  $P_\mu$  (derived in Section 2) that corresponds with a generic piecewise smooth nonlinear map, and the other system is based on the Henon map. For the piecewise linear map  $P_\mu$  we present several examples including "period 2 to period  $p$ " ( $p = 5, 11, \text{ and } 52$ ), "period 2 to quasi-periodic" and "period 2 to chaotic" bifurcation. We also present an example of a border-collision bifurcation for the map  $P_\mu$  in which no attractors but chaotic saddles are involved. The system of the plane involving the Henon map at the left side and a linear map at the right side of the border, different border-collision bifurcations are observed. We present a variety of examples. Although we do not have an exhaustive list of types of border-collision bifurcation of one-parameter families of maps under consideration, we point out that several other types of bifurcation occur. We believe this phenomenon will be seen in many applications.

In Section 4 we prove that for certain one-parameter families of piecewise smooth maps exhibit a "period 2 to period 3" border-collision bifurcation. This phenomenon persists under small perturbations of the involved maps.

In Section 5, we discuss the state of the art, and pose several questions which remain unresolved. This paper does not give a complete theory, but can be considered as initiating a bifurcation theory of piecewise smooth maps.

## 2 THE BORDER-COLLISION BIFURCATION PHENOMENON

In the bifurcation theory for maps, attention is focused on differentiable maps when one or more eigenvalues of a fixed point (or periodic point) cross the unit circle. When this occurs, the nature of the fixed point changes. For example, a fixed point attractor becomes a saddle (possibly a flip saddle) or a repellor. For border crossing fixed points, the Jacobian matrix of the fixed point generally changes discontinuously, and the fixed point can for example change from being a repellor to a saddle as  $\mu$  crosses zero.

Let  $F(\cdot, \mu) = F_\mu$  be a one-parameter family of piecewise smooth maps from the phase space  $\mathbb{R}^2$  to itself, depending smoothly on the parameter  $\mu$ , and where  $\mu$  varies in a certain interval on the real line. Let  $E_\mu$  denote a fixed point of  $F_\mu$  defined on  $-\epsilon < \mu < \epsilon$ , for some  $\epsilon > 0$ . For a general approach (which is given below) we need the concept of the "orbit index" of a periodic orbit [MY]. The orbit index is a number associated with a periodic orbit, and this number is useful in understanding the patterns of bifurcations the orbit undergoes. We say an orbit of period  $p$  is typical if its Jacobian matrix exists (that is, the Jacobian matrix of the  $p$ -th iterate of the map at a point of the orbit) and neither  $+1$  nor  $-1$  is an eigenvalue (of the Jacobian matrix). For typical orbits, the orbit index is  $-1$ ,  $0$ , or  $+1$ . The orbit index is a bifurcation invariant in the sense that if one examines the periodic orbits that collapse to the fixed point  $E_\mu$  as  $\mu \rightarrow 0$ , and adds the orbit indexes of the periodic orbits that exist just before a bifurcation, then that sum equals the corresponding sum just after that bifurcation. Suppose a typical periodic orbit  $PO$  of a map  $F$

has (minimum) period  $p$ . The orbit index of that orbit depends on the eigenvalues of the Jacobian matrix  $A_p$  of the map  $F^p$  at one of the points in  $PO$ . Now we define the orbit index  $I_{PO}$  of  $PO$ . Let  $m$  be the number of real eigenvalues of  $A_p$  smaller than  $-1$ , and let  $n$  be the number of real eigenvalues of  $A_p$  greater than  $+1$ . The orbit index  $I_{PO}$  of  $PO$  is defined by

$$I_{PO} = 0 \text{ if } m \text{ is odd;}$$

$$I_{PO} = -1 \text{ if } m \text{ is even and } n \text{ is odd;}$$

$$I_{PO} = +1 \text{ if both } m \text{ and } n \text{ are even.}$$

If the orbit index  $= -1$ , then the orbit is called a regular saddle. The typical orbits with orbit index  $+1$  in the plane are repellers and attractors and fixed points with non-real eigenvalues. The definition of orbit index is technical when a point of the orbit lies on the boundary and so does not have a Jacobian matrix, and the definition is unnecessary since we consider orbits for  $\mu \neq 0$ .

For a moment, assume that  $E_\mu$  is in the interior of the region  $R_A$  (or the region  $R_B$ ), and write  $\lambda$  and  $\nu$  for the eigenvalues of  $DF_\mu(E_\mu)$ . If neither of the two eigenvalues  $\lambda$  and  $\nu$  is on the unit circle, then the fixed point  $E_\mu$  is called a flip saddle (and has index 0) if  $\lambda < -1 < \nu < 1$ ;  $E_\mu$  is a regular saddle (and has index  $-1$ ) if  $-1 < \nu < 1 < \lambda$ ;  $E_\mu$  is a repellor (and has index  $+1$ ) if both  $|\lambda| > 1$  and  $|\nu| > 1$ ; and  $E_\mu$  is an attractor (and has index  $+1$ ) if both  $|\lambda| < 1$  and  $|\nu| < 1$ . Note that  $E_\mu$  has orbit index  $+1$  if the eigenvalues are not real. Hence, a typical fixed point is a flip saddle, a regular saddle, a repellor or an attractor. Similarly, the nature of periodic points is defined.

Now we are able to provide a definition of the notion "border-

collision bifurcation". Let the regions  $R_A$  and  $R_B$ , the map  $F_\mu$  and the fixed point (periodic point)  $E_\mu$  be as above. Assume there exists a number  $\varepsilon > 0$  such that (1)  $E_0$  is on the border of the two regions  $R_A$  and  $R_B$ , (2) for  $-\varepsilon < \mu < 0$  the fixed point  $E_\mu$  is in the region  $R_A$ , and its index is  $I_A$ , and (3) for  $0 < \mu < \varepsilon$  the fixed point  $E_\mu$  is in the region  $R_B$ , and its index is  $I_B$ . If  $I_A$  and  $I_B$  are different, then (as stated below) some bifurcation must occur at  $E_0$ , since the orbit index of  $E_\mu$  is changing from  $I_A$  to  $I_B$ , while the parameter  $\mu$  is increasing from  $-\varepsilon$  to  $+\varepsilon$ .

We say a periodic orbit PO is an isolated border crossing orbit if (1) PO includes a point that is a border crossing fixed point under some iterate of the map, and (2) the orbit PO is isolated in phase space when  $\mu = 0$ , that is, in the plane there exist a neighborhood  $U$  of the orbit PO such that PO is the only periodic orbit in  $U$  when  $\mu = 0$ . From the topological degree theory as described in [MY] (see also [AYY] for the two dimensional case), the following "Border-Collision Bifurcation" result follows after some minor modifications.

**BORDER-COLLISION BIFURCATION THEOREM.** For each two-dimensional piecewise smooth map and depending smoothly on a parameter  $\mu$ , if the index of an isolated border crossing orbit changes as  $\mu$  crosses 0, then at  $\mu = 0$  a bifurcation occurs at this point, a bifurcation involving at least one additional periodic orbit.

This result says that additional fixed points or periodic points must bifurcate from  $E_0$  at  $\mu = 0$ . These bifurcating orbits need not to be stable. An example of the preservation of orbit

index occurs with a period doubling bifurcation. If for  $\mu < 0$  there is an attracting fixed point (and no other entering orbits), the total index is +1. Then for  $\mu > 0$  we can have a flip saddle (orbit index 0) and a period 2 attractor (orbit index +1). Hence, the sum of the orbit indices before and after  $\mu = 0$  is +1. Note that the two points of the period 2 orbit are collectively assigned +1, not individually, since that orbit has index +1. Since this bifurcation occurs while the fixed point (or periodic point) collides with the border of the regions  $R_A$  and  $R_B$ , we call it a border-collision bifurcation. In other words, a border-collision bifurcation is a bifurcation at a fixed point (or periodic point) on the border of two regions when the orbit index of the fixed point (or periodic point) before the collision with the border is different from the orbit index of the fixed point after the collision.

We derive the map  $P_\mu$  that was introduced in Section 1, from nonlinear piecewise smooth maps. We assume coordinates are chosen so that the curve  $\Gamma$  is a straight line. Let  $z$  denote any vector in the plane, and write  $F_\mu(z) = F(z; \mu)$ , and write  $z_0 = E_0$ . From the assumption  $F_\mu$  is piecewise smooth, we have that on each of the regions  $R_A$  and  $R_B$

$$F(z; \mu) = F(z_0; 0) + D_z F(z_0, 0)(z - z_0) + D_\mu F(z_0, 0)\mu + \text{H.O.T.}$$

where H.O.T. stands for Higher Order Terms. Hence, there exist matrices  $M_A$  and  $M_B$  and vectors  $v_A$  and  $v_B$  such that if  $z$  is in the region  $R_A$  then

$$F(z; \mu) = F(z_0; 0) + M_A(z - z_0) + v_A\mu + \text{H.O.T.}$$

and if  $z$  is in the region  $R_B$  then

$$F(z;\mu) = F(z_0;0) + M_B(z-z_0) + v_B\mu + \text{H.O.T.}$$

Let  $e_1$  be the unit vector tangent to  $\Gamma$  at  $z_0$ . The assumption  $F_\mu$  is piecewise smooth and depends smoothly on  $\mu$  implies  $M_A e_1 = M_B e_1 = e_2$  and  $v_A = v_B = v$ . Choose coordinates so that  $z_0 = 0$ , so  $F(z_0,0) = 0$ . Assume that  $e_2$  is independent of  $e_1$ , so we may use  $e_1$  and  $e_2$  as basis vectors. We let  $e_1$  and  $e_2$  be the basis vectors of the plane. We assume that  $e_2$  is independent of  $v$  and that  $v$  is not parallel with  $e_1 - e_2$ . We claim that by change of variables and by rescaling  $\mu$  we may assume that  $v = e_1$ . Write  $v = (v_x, v_y)$ . We now assume that  $v_x \neq 0$ . We can make  $v_y = 0$  after a change of variables, and  $v_x = 1$  by rescaling of  $\mu$ . If  $v_y$  is not 0 then we can change variables, setting  $\bar{y} = y - v_y \mu$  (where  $x$  is unchanged), and the new vector  $v$  for the  $(x, \bar{y})$  system will have its second coordinate 0. By rescaling  $\mu$ , that is, by introducing  $\bar{\mu} = \mu v_x$ , we can change the system so that the new vector  $v$  is  $(1,0)$ , when  $\bar{\mu}$  is the parameter. Therefore, we may write  $M_A = \begin{bmatrix} a & 1 \\ b & 0 \end{bmatrix}$ ,  $M_B = \begin{bmatrix} c & 1 \\ d & 0 \end{bmatrix}$ , and  $v = (1,0)$ . Since all these assumptions are generic, we say the prototype piecewise linear form of  $F_\mu$  for  $\mu$  small is defined by

$$F(z;\mu) = \begin{bmatrix} a & 1 \\ b & 0 \end{bmatrix} z + \mu(1,0) \text{ if } z \text{ is in the region } R_A,$$

$$F(z;\mu) = \begin{bmatrix} c & 1 \\ d & 0 \end{bmatrix} z + \mu(1,0) \text{ if } z \text{ is in the region } R_B.$$

To write the prototype piecewise linear form of  $F_\mu$  in the form of the map  $P_\mu$ , let  $u = (\frac{a+c}{2}, \frac{b+d}{2})$ , and  $w = (\frac{a-c}{2}, \frac{b-d}{2})$ .

We observe the following fact. Assume that the fixed point  $E_\mu$  is a flip saddle (orbit index 0) in region  $R_A$  and a repellor with complex eigenvalues (orbit index +1) in region  $R_B$ . If there exists

a stable periodic orbit with period 2 in  $R_A$  that converges to  $E_0$  when  $u$  approaches 0, then the total degree in  $R_A$  is +1. Hence, if there exists a stable periodic orbit in  $R_B$  that converges to  $E_0$  when  $\mu$  goes to 0, then there must exist a regular saddle periodic orbit of the same period (orbit index -1) in  $R_B$  that converges to  $E_0$  when  $\mu$  goes to 0, since the total orbit index is a bifurcation invariant. Consequently, for the family of maps  $f_\mu$  in the Section 1 exhibiting a "period two to period three" bifurcation in figure 2, there must also exist a regular saddle periodic orbit with period 3.

#### PERIOD TWO TO PERIOD THREE BORDER-COLLISION BIFURCATION THEOREM.

Let  $F_\mu$  be a one-parameter family of piecewise smooth maps which has a prototype piecewise linear form at  $\mu = 0$ , and assume that (1)  $a < -1$ ,  $c < -1$ ,  $d < -1$ ; (2)  $c^2 + 4d < 0$ ; and (3)  $0 < a(ac + d) < 1$ . Then, there exists  $\epsilon > 0$  such that if  $|b| < \epsilon$ , then the family  $F_\mu$  has a "period two to period three" border-collision bifurcation at  $(0,0)$ .

We point out that the border-collision bifurcations persist under small perturbations. The proof follows of the Theorem from the result obtained in Section 4. The geometrical proof given in Section 4, might give insight why other bifurcations (for example, period 5 to period 2 bifurcation) may occur in piecewise smooth systems. Presumably, the method of proof only works if one of the two maps involved has a small Jacobian. Hence, when the piecewise smooth map consists of maps that all have Jacobian bounded (far) away from zero, new techniques have to be developed to obtain

rigorous border-collision bifurcation results.

### 3. A VARIETY OF BORDER-COLLISION BIFURCATIONS.

In this Section we present a variety of numerical examples exhibiting a border-collision bifurcation. The first series of examples is from the piecewise linear map  $P_\mu$ , and the second series is based on the Hénon map. We will present examples showing that in a border-collision bifurcation not only attracting periodic orbits are involved, but also chaotic saddles may play a role. Therefore, in order to describe the qualitatively different border-collision bifurcations in a consistent manner, we refer to the invariant sets that are involved in the border-collision bifurcation. A chaotic saddle is a compact, invariant set that is not an attractor which contains a chaotic trajectory. If an attractor  $A$  of a map  $F$  is an attracting periodic orbit with period  $p$ , then we call  $A$  a period  $p$  attractor, and we say instead of "period two to period three" bifurcation a bifurcation from a period 2 attractor to a period 3 attractor.

The bifurcation diagrams below show the long term behavior of the coordinate  $x$  for  $\mu$  between  $-0.1$  and  $0.2$ . The diagrams have been constructed as follows. For the minimum value  $-0.1$  of  $\mu$ , and initial value  $(0,0)$ , calculate the first 200 points (transient time 200) of the orbit and plot the next 1000 points of the orbit. Increase  $\mu$  slightly, say by  $0.001$ , take for the initial value the last point which was plotted, calculate 200 points of this orbit and plot the next 1000 points. Increase  $\mu$  again, and continue increasing until  $\mu$  achieves the maximum value  $0.2$ . Hence, once the orbit is close to an attractor, as the parameter is increased,

this attractor is "followed" as long as it exists. In the diagrams, the x-coordinate is plotted vertically, and the parameter  $\mu$  is plotted horizontally.

Define the map  $GL_\mu$  from the plane to itself to be the prototype piecewise linear form of  $F_\mu$ , that is,

$$GL_\mu(x,y) = (ax + y, bx) + \mu(1,0) \quad \text{if } x \leq 0$$

$$GL_\mu(x,y) = (cx + y, dx) + \mu(1,0) \quad \text{if } x \geq 0$$

Recall that the map  $GL_\mu$  is equivalent to the map  $P_\mu$ , since to write the map  $GL_\mu$  in the form of the map  $P_\mu$ , let  $u = (\frac{a+c}{2}, \frac{b+d}{2})$ , and  $w = (\frac{a-c}{2}, \frac{b-d}{2})$ . We present a few numerical examples for this map  $GL_\mu$  exhibiting a border-collision bifurcation. In all these examples, the fixed point is a flip saddle for  $\mu < 0$  and a repellor with complex eigenvalues for  $\mu > 0$ .

EXAMPLE 1. The presumably simplest border-collision bifurcation is from a period 2 attractor to a period 3 attractor presented in Figure 1. We present parameter values for which the map  $GL_\mu$  shows a bifurcation from a period 2 attractor to a period p attractor for a variety of period p.

For  $a = -1.25$ ,  $b = -.035$ ,  $c = -2$ ,  $d = -1.75$ , the bifurcation diagram in Figure 3a exhibits a bifurcation from a period 2 attractor to a period 5 attractor.

For  $a = -1.25$ ,  $b = -0.0435$ ,  $c = -2$ ,  $d = -2.175$ , the bifurcation diagram in Figure 3b exhibits a bifurcation from a period 2 attractor to a period 11 attractor.

For  $a = -1.25$ ,  $b = -0.03943$ ,  $c = -2$ ,  $d = -1.9715$ , the bifurcation diagram in Figure 3c exhibits a bifurcation from a period 2 attractor to a period 52 attractor.

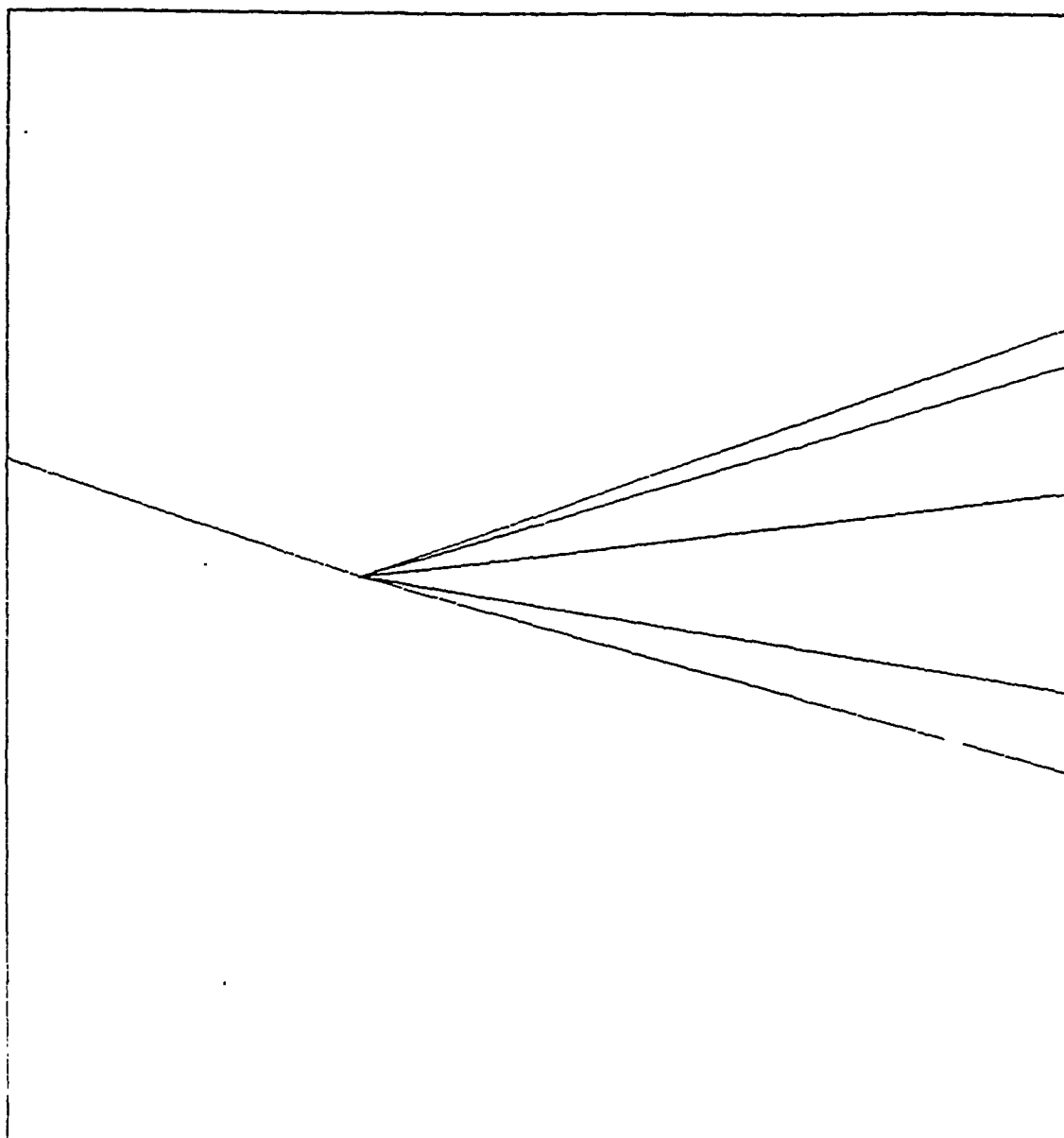


Figure 3a.

Bifurcation diagram of  
 $GL_{\mu}(x,y) = (-1.25x + y + \mu, -0.035x)$  if  $x \leq 0$ , and  
 $= (-2x + y + \mu, -1.75x)$  if  $x \geq 0$   
 exhibits at  $\mu_0 = 0$  a border-collision bifurcation from a period 2  
 attractor to a period 5 attractor. The parameter  $\mu$  (plotted  
 horizontally) varies from  $-0.1$  to  $0.2$ , and the coordinate  $x$  is  
 plotted vertically,  $-0.3 \leq x \leq 0.3$ .

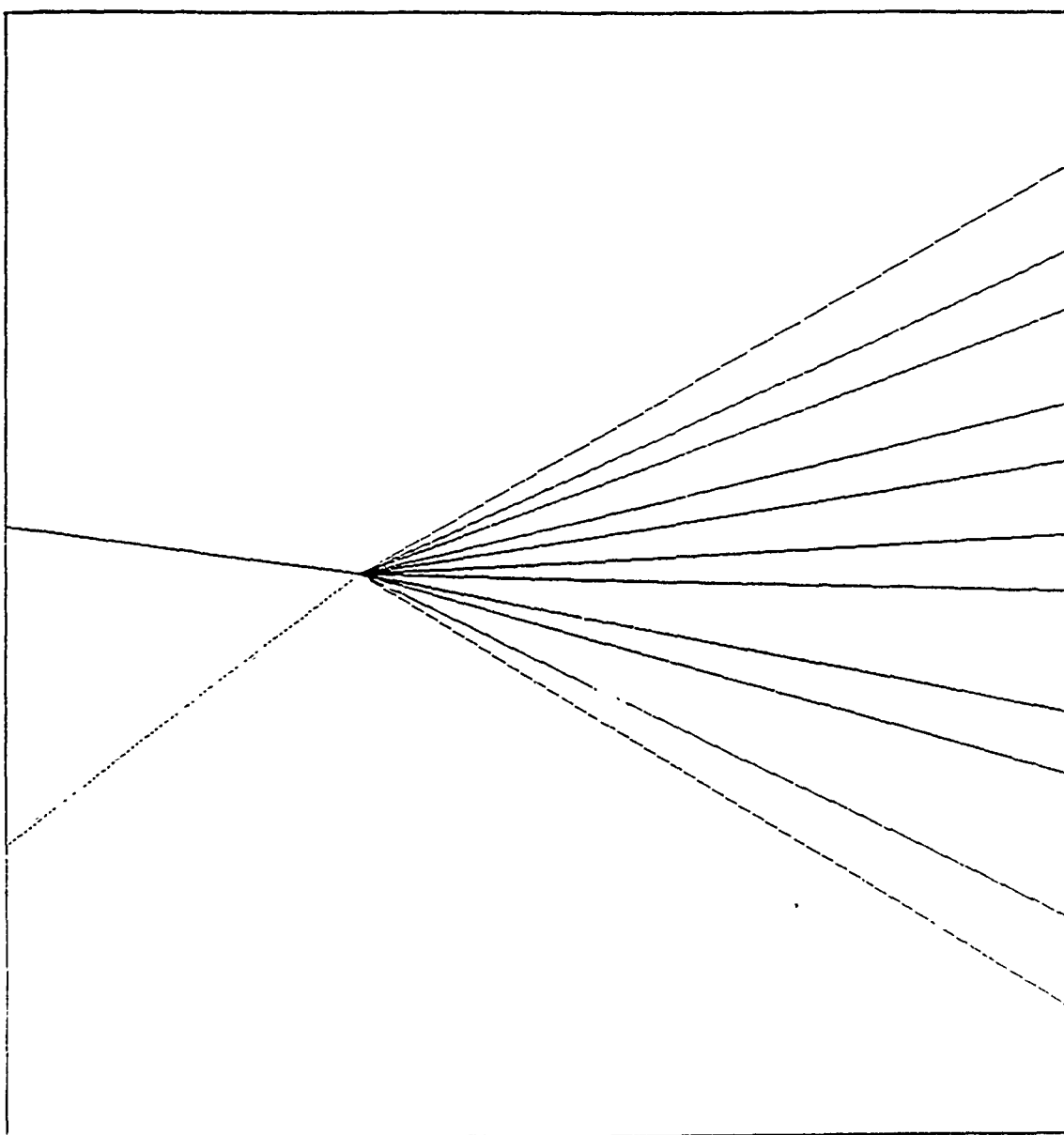


Figure 3b.

Bifurcation diagram of

$$GL_{\mu}(x,y) = (-1.25x + y + \mu, -0.0435x) \text{ if } x \leq 0, \text{ and} \\ = (-2x + y + \mu, -2.175x) \text{ if } x \geq 0$$

exhibits at  $\mu_0 = 0$  a border-collision bifurcation from a period 2 attractor to a period 11 attractor. The parameter  $\mu$  (plotted horizontally) varies from -0.1 to 0.2, and the coordinate  $x$  is plotted vertically,  $-0.3 \leq x \leq 0.3$ .

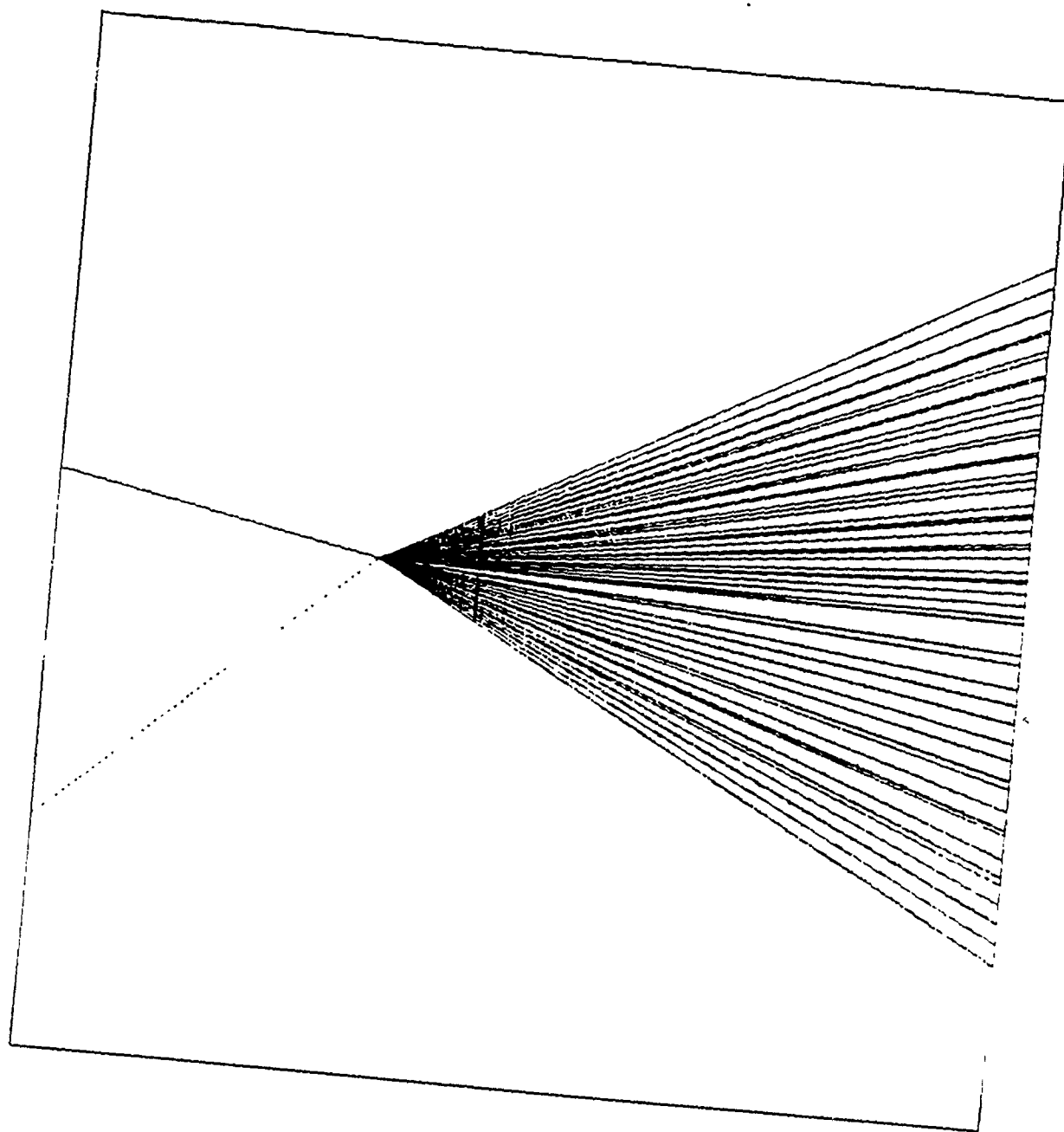


Figure 3c.

Bifurcation diagram of  
 $GL_{\mu}(x,y) = (-1.25x + y + \mu, -0.03943x)$  if  $x \leq 0$ , and  
 $= (-2x + y + \mu, -1.9715x)$  if  $x \geq 0$   
 exhibits at  $\mu_0 = 0$  a border-collision bifurcation from a period 2  
 attractor to a period 52 attractor. The parameter  $\mu$  (plotted  
 horizontally) varies from  $-0.1$  to  $0.2$ , and the coordinate  $x$  is  
 plotted vertically,  $-0.3 \leq x \leq 0.3$ .

For other choices for  $a$ ,  $b$ ,  $c$ , and  $d$  we have found bifurcations from a period 2 attractor to a period  $p$  attractor, where  $p = 6, 7, 8, 9, 10, 11, 13, 19, 21, 23, 29, 31, 37, 41$ , etc.

EXAMPLE 2. The simplest border-collision bifurcation in which chaotic attractors are involved is presumably the bifurcation from a period 2 attractor to a (1-piece) chaotic attractor. Frequently, the border-collision bifurcation from a period 2 attractor to a  $p$ -piece chaotic attractor is observed.

For  $a = -1.25$ ,  $b = -0.042$ ,  $c = -2$ , and  $d = -2.1$ , the bifurcation diagram in Figure 4a exhibits a bifurcation from a period 2 attractor to a 1-piece chaotic attractor.

For  $a = -1.36$ ,  $b = -0.12$ ,  $c = -2$ , and  $d = -2$ , the bifurcation diagram in Figure 4b seems to exhibit a bifurcation from a period 2 attractor to a 12-piece chaotic attractor, but using the phase space it turns out that the bifurcation is from a period 2 attractor to a 18-piece chaotic attractor.

We have observed many other values of  $p$ , the map  $GL_\mu$  shows a bifurcation from period 2 attractor to  $p$ -piece chaotic attractor.

For the selection  $a = -1.25$ ,  $b = -0.03865$ ,  $c = -2$ , and  $d = -1.9325$ , we obtain a bifurcation diagram similar to figure 4a, but in this case the border-collision bifurcation is from a period 2 attractor to a what appears to be quasi-periodic attractor.

EXAMPLE 3. A border-collision bifurcation in which chaotic saddles (rather than attractors) are involved, will not be exhibited by bifurcation diagrams. Therefore, some other numerical method is needed to detect these sets. We use the Saddle Straddle

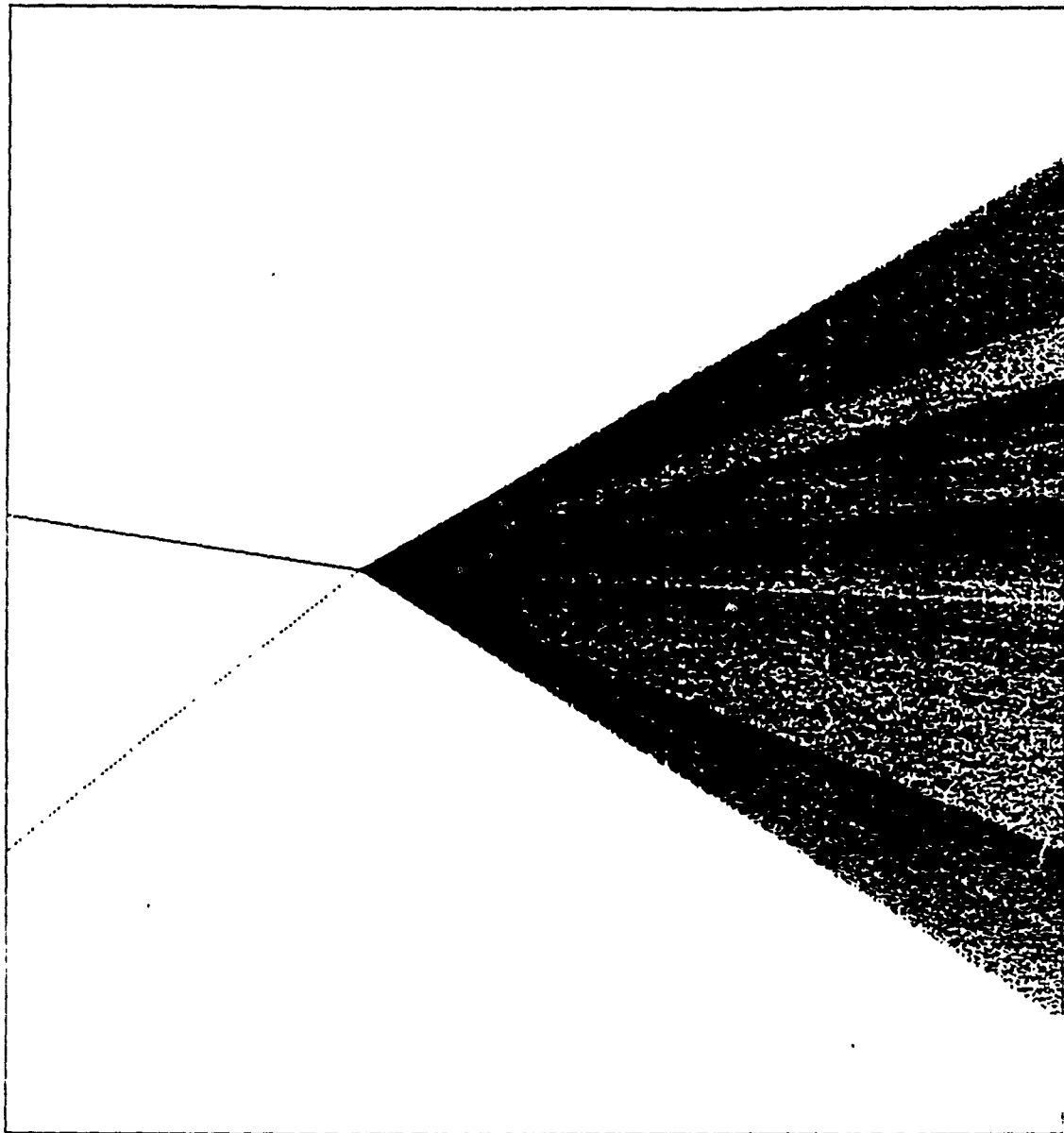


Figure 4a.

Bifurcation diagram of  
 $GL_{\mu}(x,y) = (-1.25x + y + \mu, -0.042x)$  if  $x \leq 0$ , and  
 $= (-2x + y + \mu, -2.1x)$  if  $x \geq 0$   
 exhibits at  $\mu_0 = 0$  a border-collision bifurcation from a period 2  
 attractor to a 1-piece chaotic attractor. The parameter  $\mu$  (plotted  
 horizontally) varies from -0.1 to 0.2; the coordinate  $x$  is plotted  
 vertically,  $-0.3 \leq x \leq 0.3$ .

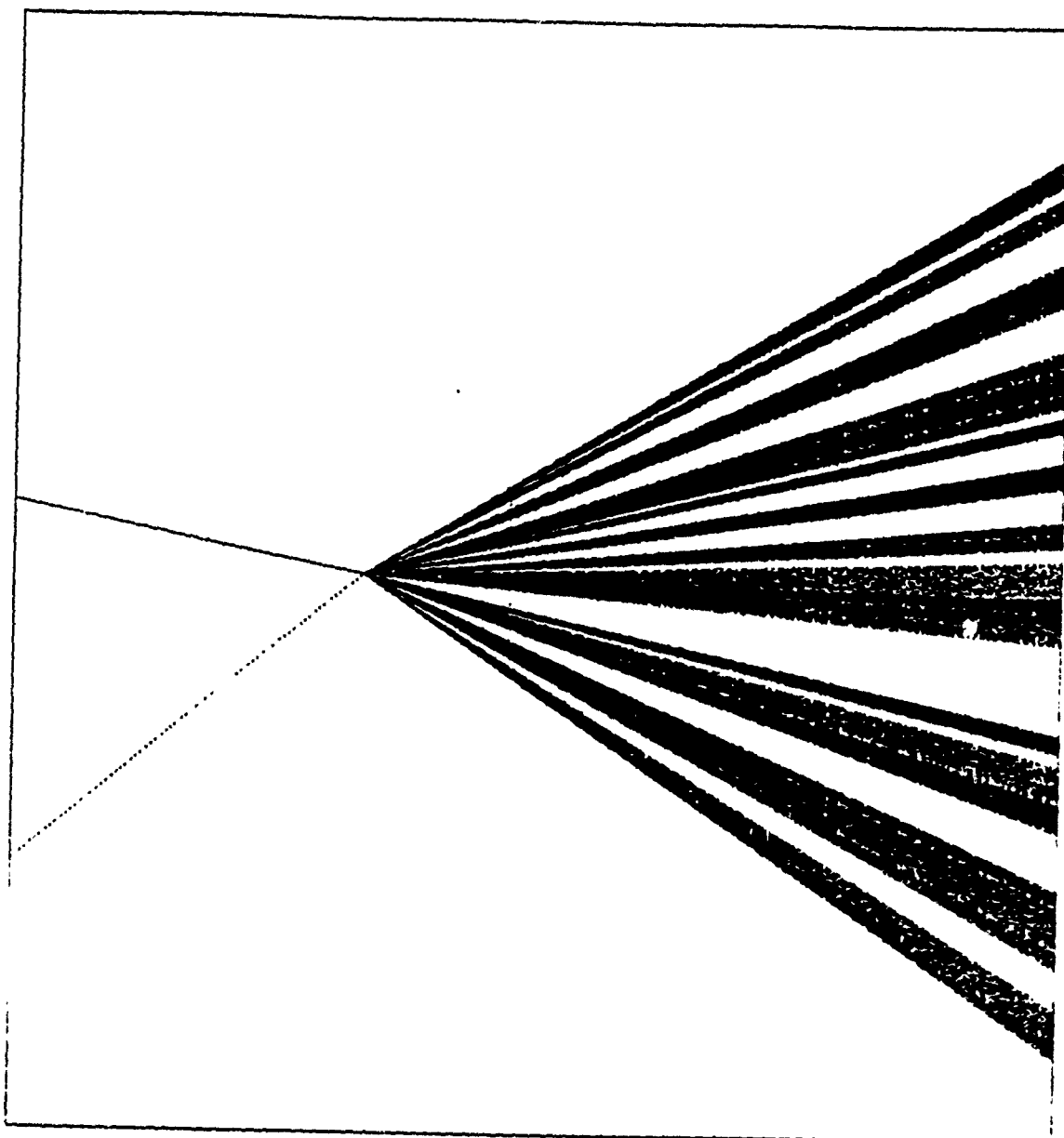


Figure 4b.

Bifurcation diagram of

$$GL_{\mu}(x,y) = (-1.36x + y + \mu, -0.12x) \text{ if } x \leq 0, \text{ and} \\ = (-2x + y + \mu, -2x) \text{ if } x \geq 0$$

exhibits at  $\mu_0 = 0$  a border-collision bifurcation from a period 2 attractor to a 18-piece chaotic attractor. The parameter  $\mu$  (plotted horizontally) varies from -0.1 to 0.2; the coordinate  $x$  is plotted vertically,  $-0.3 \leq x \leq 0.3$ .

Trajectory (SST) method introduced in [NY] to detect such sets.

We select  $a = -1.25$ ,  $b = 0.18$ ,  $c = 2$ , and  $d = -3$ . For  $\mu = -0.05$  the invariant set (obtained by the SST method) is presented in Figure 5a, and the invariant set for  $\mu = 0.05$  is in Figure 5b. Presumably, it is correct to say that the border-collision bifurcation is a bifurcation from a chaotic saddle to another chaotic saddle.

Now we present a few examples based on the Henon map. In fact, in these examples we have a moving border. Define the map  $H$  from the plane to itself by

$$H(x,y) = (A - x^2 + By, x)$$

and define the map  $L_\mu$  ( $-\infty < \mu < \infty$ ) from the plane to itself by

$$L_\mu(x,y) = (A + Cx + By - (\mu+C)\mu, Dx + (1-D)\mu)$$

The regions  $R_A$  and  $R_B$  are the half planes to the left and the right of the straight line  $x = \mu$ . The map we are investigating is defined being the Henon map on  $R_A$  and the "linear" map  $L_\mu$  on  $R_B$ . Define the one-parameter family of maps  $F_\mu$  from the plane to itself by

$$F_\mu(x,y) = \begin{cases} H(x,y) & \text{if } x \leq \mu \\ L_\mu(x,y) & \text{if } x \geq \mu \end{cases}$$

Notice that the map is smooth in each of the half planes  $x \leq \mu$  and  $x \geq \mu$ , and the line  $x = \mu$  is the border which is a smooth curve. Since the map  $F_\mu$  is continuous, it is a piecewise smooth map. Note that for this family  $F_\mu$  border-collision bifurcations occur presumably for values  $\mu_0$  different from zero.

rho = -0.0500000000

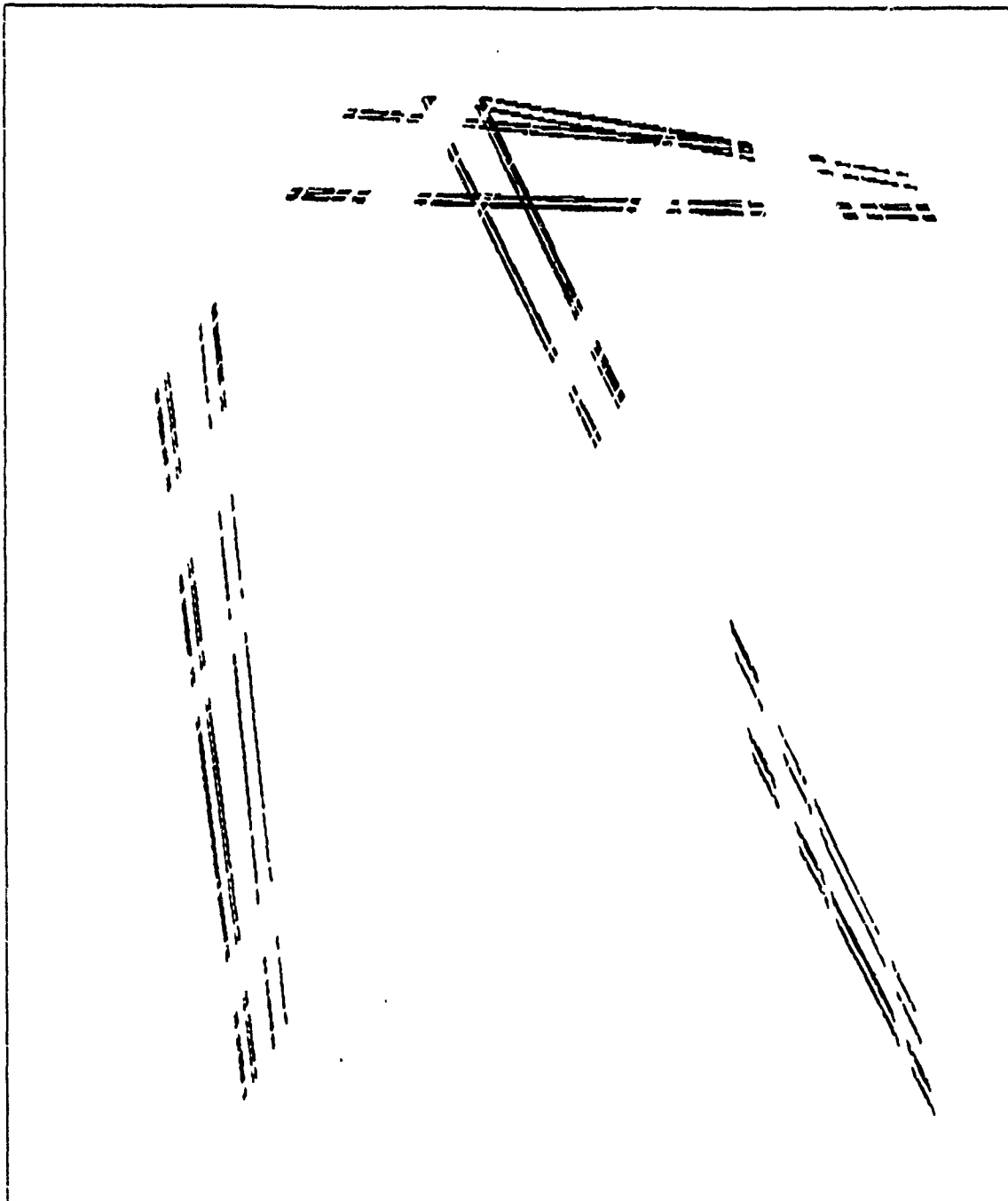


Figure 5a.

Chaotic saddle of  
 $GL_{\mu}(x,y) = (-1.25x + y + \mu, 0.18x)$  if  $x \leq 0$ , and  
 $= (2x + y + \mu, -3x)$  if  $x \geq 0$  when  $\mu = -0.05$ .

The coordinate  $x$  ( $-0.2 \leq x \leq 0.1$ ) is plotted horizontally, and the coordinate  $y$  ( $-0.25 \leq y \leq 0.02$ ) is plotted vertically.

rho = 0.0500000000

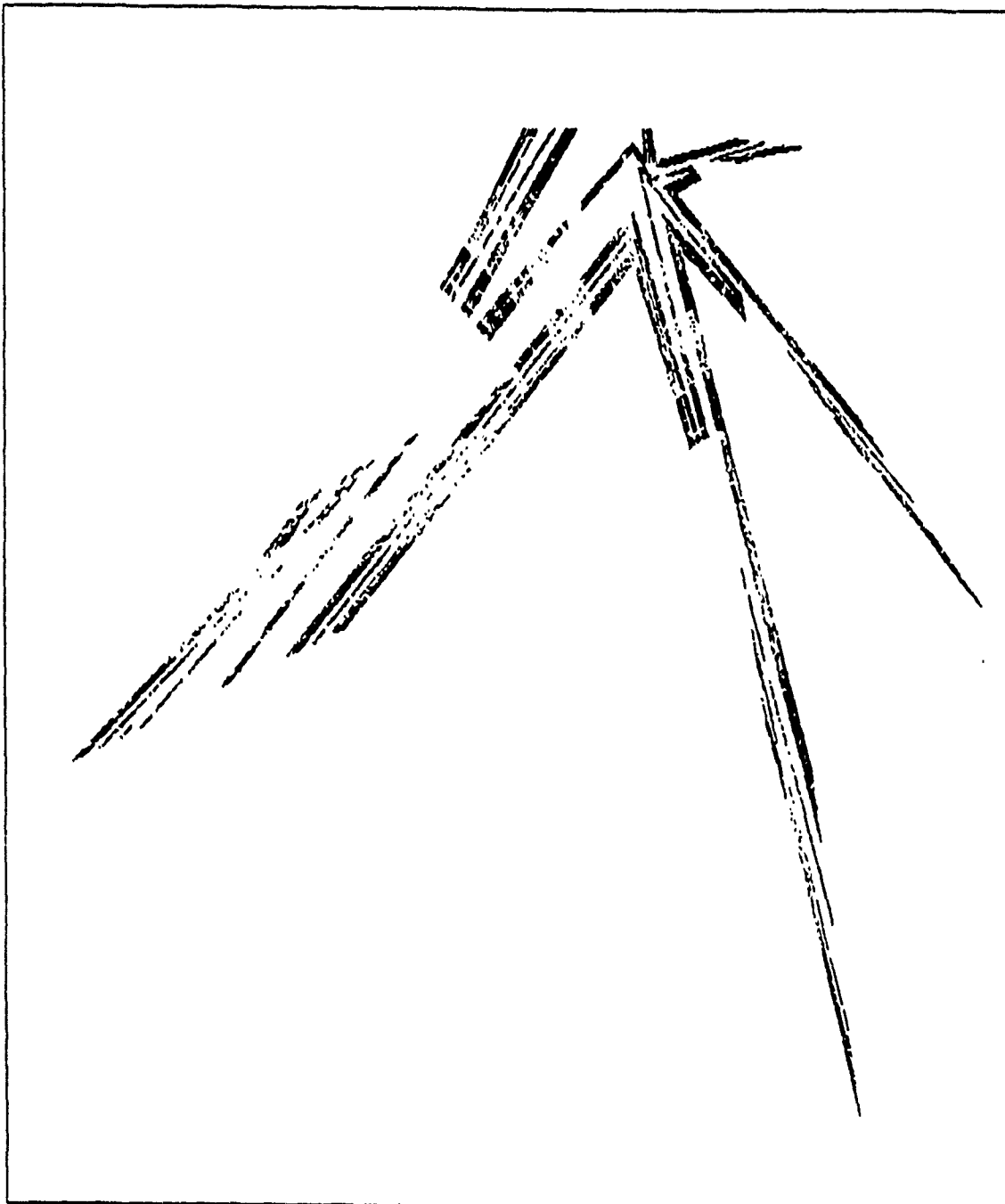


Figure 5b.

Chaotic saddle of

$$GL_{\mu}(x,y) = (-1.25x + y + \mu, 0.18x) \text{ if } x \leq 0, \text{ and} \\ = (2x + y + \mu, -3x) \text{ if } x \geq 0 \text{ when } \mu = 0.05.$$

The coordinate  $x$  ( $-1 \leq x \leq 0.6$ ) is plotted horizontally, and the coordinate  $y$  ( $-1.8 \leq y \leq 0.2$ ) is plotted vertically.

EXAMPLE 4. Simple border-collision bifurcations are bifurcations from a period  $p$  attractor to a period  $q$  attractor.

For  $A = 1.4$ ,  $B = 0.3$ ,  $C = 0.9$ , and  $D = -5$ , the bifurcation diagram in Figure 6a exhibits a bifurcation from a period 3 attractor to a period 4 attractor, where  $\mu$  (plotted horizontally) varies from 0.89 to 0.87. In the region  $R_A$  the fixed point is a flip saddle and in the region  $R_B$  the fixed point is a repellor. The border-collision bifurcation occurs at  $\mu = \mu_0 \approx 0.884$ . For  $\mu > \mu_0$  (the side of the period 3 attractor which has orbit index +1) the fixed point is a flip saddle (orbit index 0) and we find no other periodic orbits on this side of the bifurcation. For  $\mu < \mu_0$  (the side of the period 4 attractor which has orbit index +1) the fixed point is a repellor (orbit index +1); there also exists a period 4 regular saddle (orbit index -1). The regular saddle also shrinks to a point (the fixed point) as  $\mu \rightarrow \mu_0$ . Hence, the orbit index is +1 on both sides of  $\mu_0$ .

For  $A = 1.4$ ,  $B = 0.3$ ,  $C = 1$ , and  $D = -5$ , the bifurcation diagram in Figure 6b exhibits a bifurcation from a period 6 attractor to a period 4 attractor, where  $\mu$  (plotted horizontally) varies from 1.05 to 0.8. In the figure one might first notice a bifurcation from a 6-piece chaotic attractor to a period 4 attractor, but closer examination gives the above mentioned bifurcation from a period 6 attractor to a period 4 attractor. Similarly as above, the periodic orbits involved in the border-collision bifurcation that occurs at  $\mu = \mu_0 \approx 0.884$  are the following. For  $\mu > \mu_0$  there is period 6 attractor and the fixed point is a flip saddle, and for  $\mu < \mu_0$  the fixed point is a repellor and there is a period 4 attractor a period 4 regular

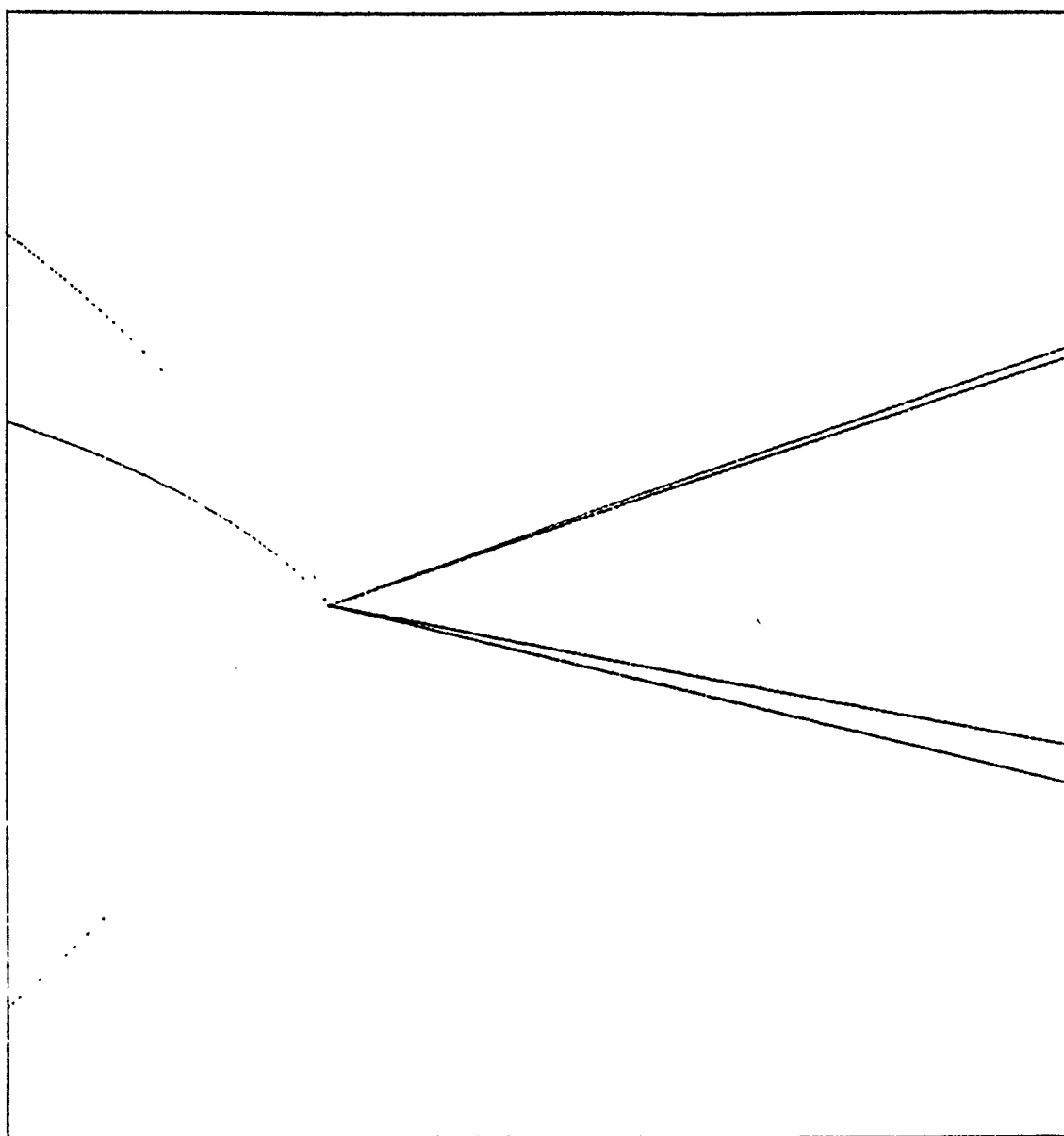


Figure 6a.

Bifurcation diagram of  
 $F_{\mu}(x,y) = (1.4 - x^2 + 0.3y, x)$  if  $x \leq \mu$ , and  
 $= (1.4 + 0.9x + 0.3y - (\mu+0.9)\mu, -5x + 6\mu)$  if  $x \geq \mu$ ,  
 exhibits at  $\mu_0 \approx 0.884$  a border-collision bifurcation from a  
 period 3 attractor to a period 4 attractor. The parameter  $\mu$   
 (plotted horizontally) varies from 0.89 to 0.87; the coordinate  $x$   
 is plotted vertically,  $0.6 \leq x \leq 1.2$ .

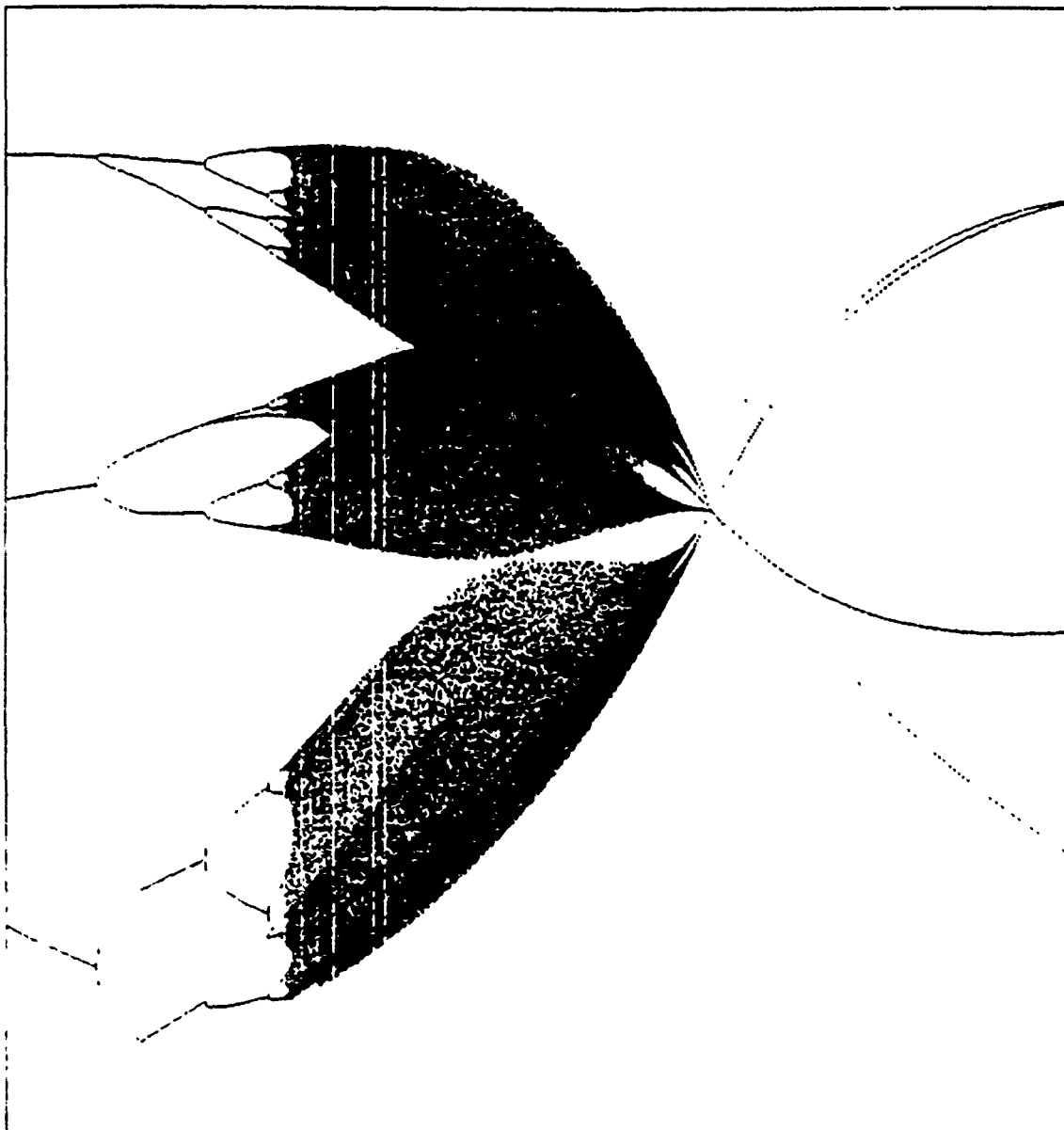


Figure 6b.

Bifurcation diagram of  
 $F_{\mu}(x,y) = (1.4 - x^2 + 0.3y, x)$  if  $x \leq \mu$ , and  
 $= (1.4 + x + 0.3y - (\mu+1)\mu, -5x + 6\mu)$  if  $x \geq \mu$ ,  
 exhibits at  $\mu_0 \approx 0.884$  a border-collision bifurcation from a  
 period 6 attractor to a period 4 attractor. The parameter  $\mu$   
 (plotted horizontally) varies from 1.05 to 0.8; the coordinate  $x$   
 is plotted vertically,  $-0.5 \leq x \leq 2$ .

saddle. Hence, the orbit index is +1 on both sides of  $\mu_0$ .

EXAMPLE 5. In this example we present two cases of a border-collision bifurcation from a period  $p$  attractor to a  $q$ -piece chaotic attractor.

For  $A = 1.4$ ,  $B = 0.3$ ,  $C = 1.1$ , and  $D = -5$ , the bifurcation diagram in Figure 7a exhibits a bifurcation from a 1-piece chaotic attractor to a period 4 attractor, where  $\mu$  (plotted horizontally) varies from 1.05 to 0.8. The border-collision bifurcation occurs at  $\mu = \mu_0 \approx 0.885$ . For  $\mu > \mu_0$  (the side with the chaotic attractor) we do not know the (total) orbit index since the chaotic attractor contains a lot of periodic orbits. For  $\mu > \mu_0$  the fixed point is a flip saddle (orbit index 0). For  $\mu < \mu_0$  (the side of the period 4 attractor which has orbit index +1) the fixed point is a repellor (orbit index +1) there also exists a period 4 regular saddle (orbit index -1). The regular saddle also shrinks to the fixed point as  $\mu \rightarrow \mu_0$ . Hence, presumably we have a border-collision bifurcation from a period 4 attractor to a 1-piece chaotic attractor.

For  $A = 1.4$ ,  $B = 0.3$ ,  $C = 1.5$ , and  $D = -4$ , the bifurcation diagram in Figure 7b exhibits a bifurcation from a 8-piece chaotic attractor to a period 5 attractor, where  $\mu$  (plotted horizontally) varies from 0.91 to 0.86. The border-collision bifurcation occurs at  $\mu = \mu_0 \approx 0.884$ . For  $\mu > \mu_0$  (the side of the 8-piece chaotic attractor) we do not know the (total) orbit index since the chaotic attractor contains a lot of periodic orbits, and the fixed point is a flip saddle (orbit index 0). For  $\mu < \mu_0$  (the side of the period 5 attractor which has orbit index +1) the fixed point

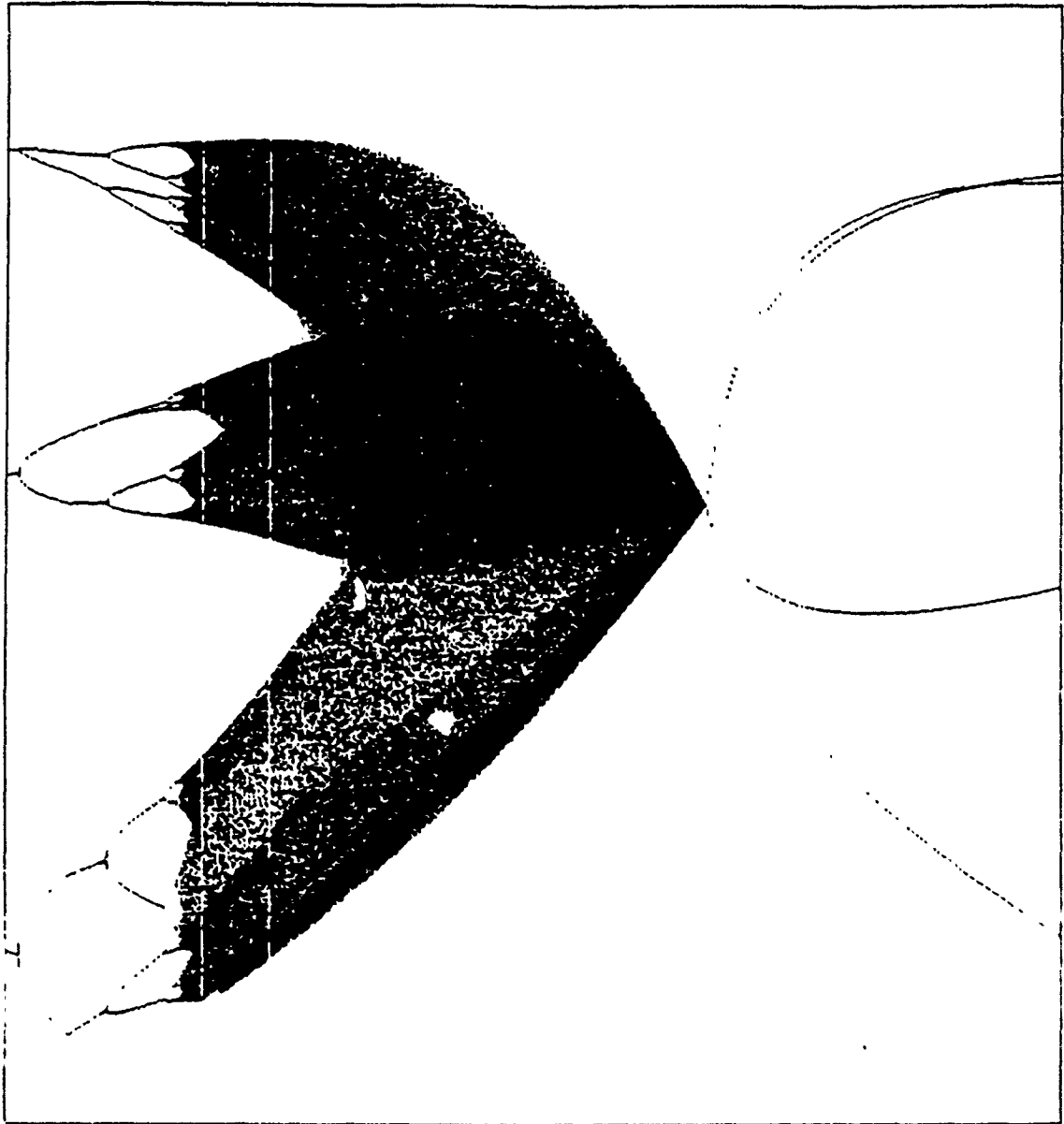


Figure 7a.

Bifurcation diagram of  
 $F_{\mu}(x, y) = (1.4 - x^2 + 0.3y, x)$  if  $x \leq \mu$ , and  
 $= (1.4 + 1.1x + 0.3y - (\mu + 1.1)\mu, -5x + 6\mu)$  if  $x \geq \mu$ ,  
 exhibits at  $\mu_0 \approx 0.885$  a border-collision bifurcation from a  
 1-piece chaotic attractor to a period 4 attractor. The parameter  $\mu$   
 (plotted horizontally) varies from 1.05 to 0.8; the coordinate  $x$   
 is plotted vertically,  $-0.5 \leq x \leq 2$ .

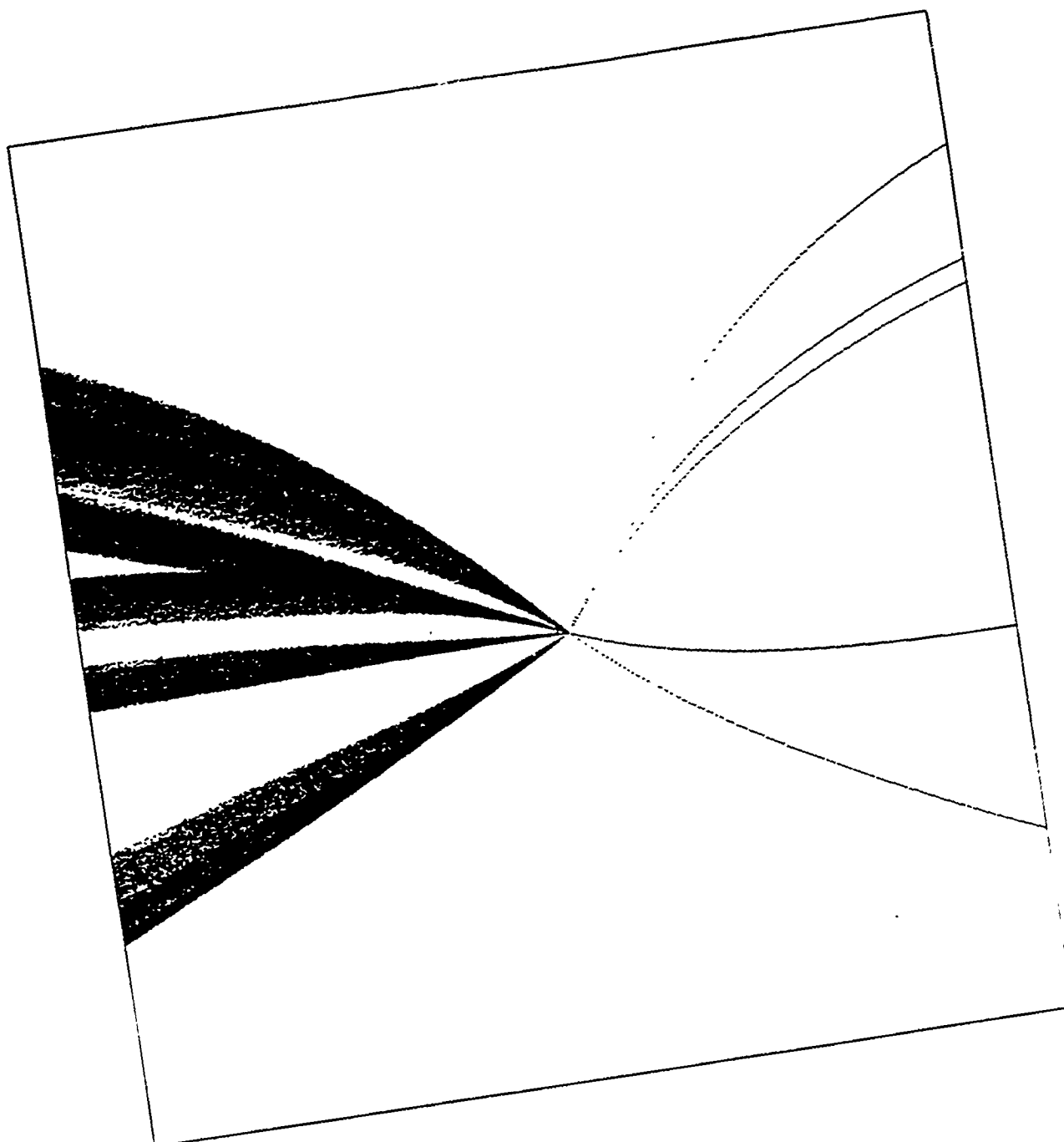


Figure 7b.

Bifurcation diagram of  $F_\mu(x, y) = (1.4 - x^2 + 0.3y, x)$  if  $x \leq \mu$ , and  $F_\mu(x, y) = (1.4 + 1.5x + 0.3y - (\mu + 1.5)\mu, -4x + 5\mu)$  if  $x \geq \mu$ , exhibits at  $\mu_0 \approx 0.884$  a border-collision bifurcation from a 8-piece chaotic attractor to a period 5 attractor. The parameter  $\mu$  (plotted horizontally) varies from 0.91 to 0.86; the coordinate  $x$  is plotted vertically,  $0 \leq x \leq 2$ .

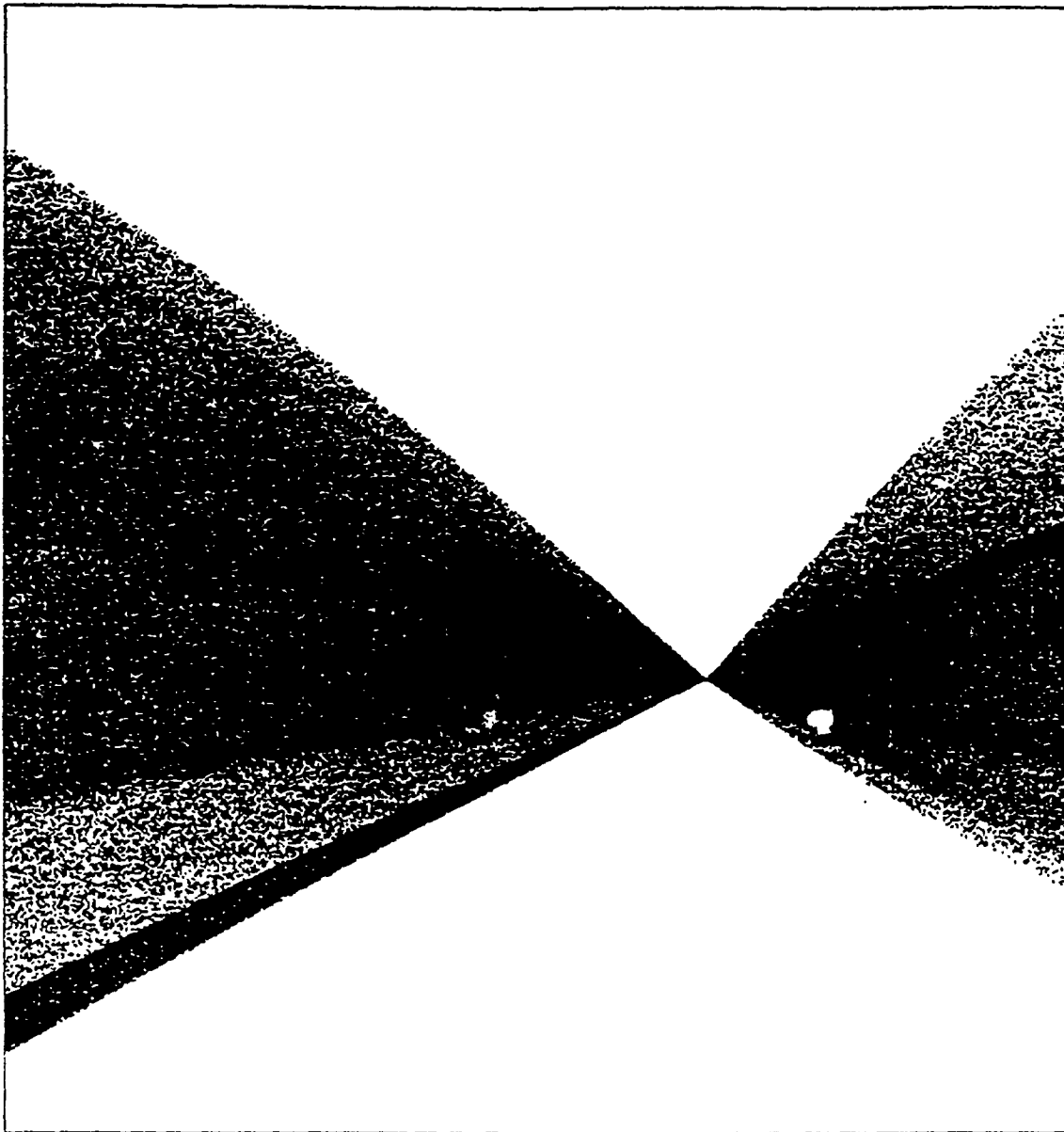


Figure 8.

Bifurcation diagram of  
 $F_{\mu}(x,y) = (1.4 - x^2 + 0.3y, x)$  if  $x \leq \mu$ ; and  
 $= (1.4 + 1.2x + 0.3y - (\mu+1.2)\mu, -4x + 5\mu)$  if  $x \geq \mu$ ,  
 exhibits at  $\mu_0 \approx 0.884$  a border-collision bifurcation from a  
 1-piece chaotic attractor to a 1-piece chaotic attractor. The  
 parameter  $\mu$  (plotted horizontally) varies from 0.95 to 0.85; the  
 coordinate  $x$  is plotted vertically,  $0.4 \leq x \leq 1.6$ .

is a repeller (orbit index +1); there also exists a period 5 regular saddle (orbit index -1). The regular saddle also shrinks to the fixed point as  $\mu \rightarrow \mu_0$ . In the figure one might first notice a bifurcation from a 5-piece chaotic attractor to a period 5 attractor, but closer examination in the phase space gives the above mentioned bifurcation from a 8-piece chaotic attractor to a period 5 attractor. Hence, presumably we have a border-collision bifurcation from a period 5 attractor to a 8-piece chaotic attractor.

EXAMPLE 6. Border-collision bifurcation from a  $p$ -piece chaotic attractor to a  $q$ -piece chaotic attractor. We present just one example, namely  $p = \alpha = 1$ .

For  $A = 1.4$ ,  $B = 0.3$ ,  $C = 1.2$ , and  $D = -4$ , the bifurcation diagram in Figure 8 exhibits a bifurcation from a 1-piece chaotic attractor to a 1-piece chaotic attractor, where  $\mu$  (plotted horizontally) varies from 0.95 to 0.85. The border-collision bifurcation occurs at  $\mu = \mu_0 \approx 0.884$  and we only can say that on both sides infinitely many periodic orbits are involved in the border-collision bifurcation, since the attractors are chaotic. Hence, presumably we have a border-collision bifurcation from a 1-piece chaotic attractor to a 1-piece chaotic attractor.

EXAMPLE 7. In this example we show that coexisting attractors of different nature can be involved on the same side of a border-collision bifurcation.

For  $A = 1.4$ ,  $B = 0.3$ ,  $C = 1.4$ , and  $D = -4$ , the bifurcation diagram in Figure 9a exhibits a bifurcation from a 5-piece chaotic

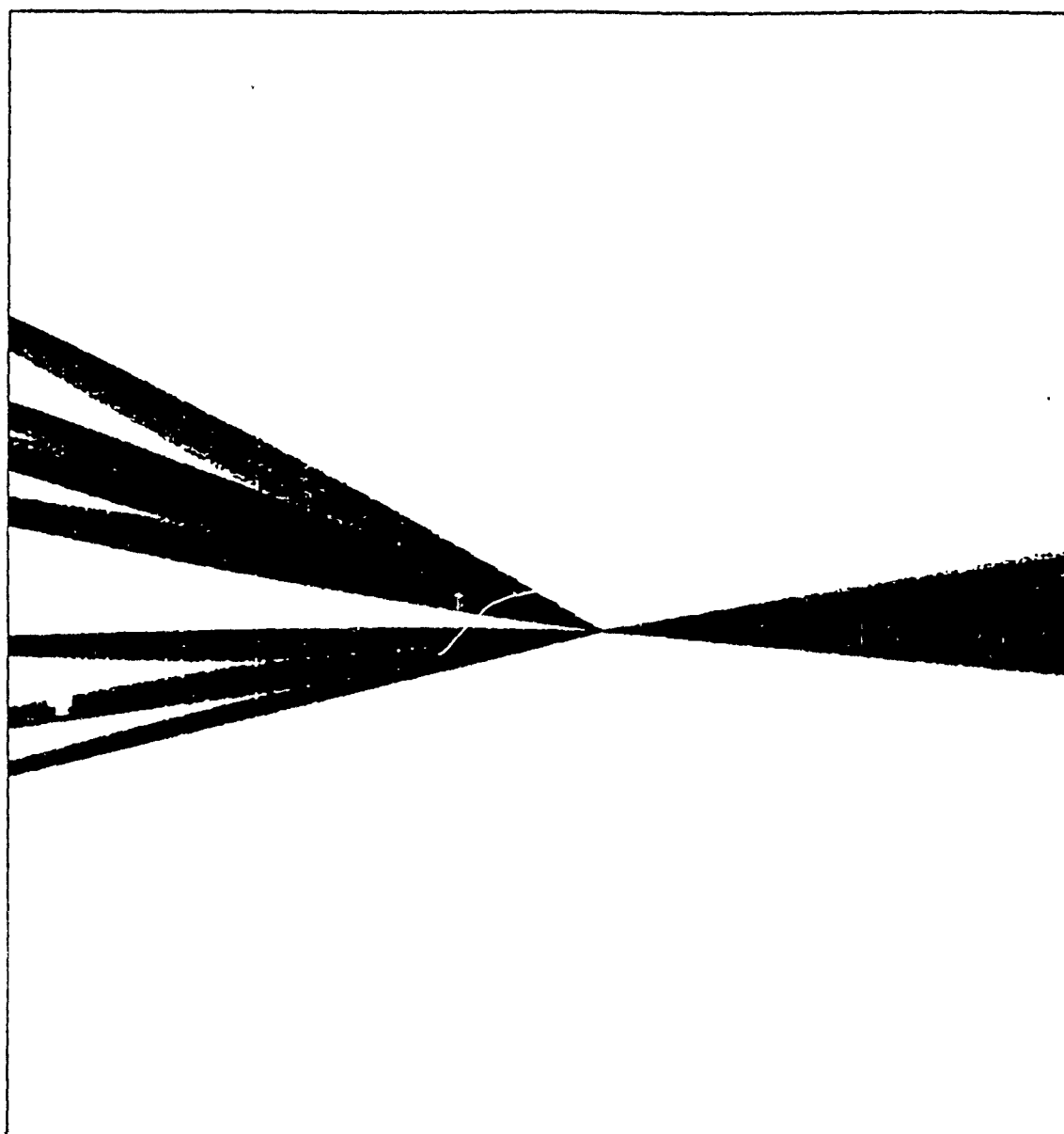


Figure 9a.

Bifurcation diagram of  
 $F_\mu(x, y) = (1.4 - x^2 + 0.3y, x)$  if  $x \leq \mu$ , and  
 $= (1.4 + 1.4x + 0.3y - (\mu + 1.4)\mu, -4x + 5\mu)$  if  $x \geq \mu$ ,  
 exhibits at  $\mu_0 \approx 0.884$  a border-collision bifurcation from a  
 5-piece chaotic attractor to a 1-piece chaotic attractor. The  
 parameter  $\mu$  (plotted horizontally) varies from 0.87 to 0.895; the  
 coordinate  $x$  is plotted vertically,  $0.3 \leq x \leq 1.6$ .

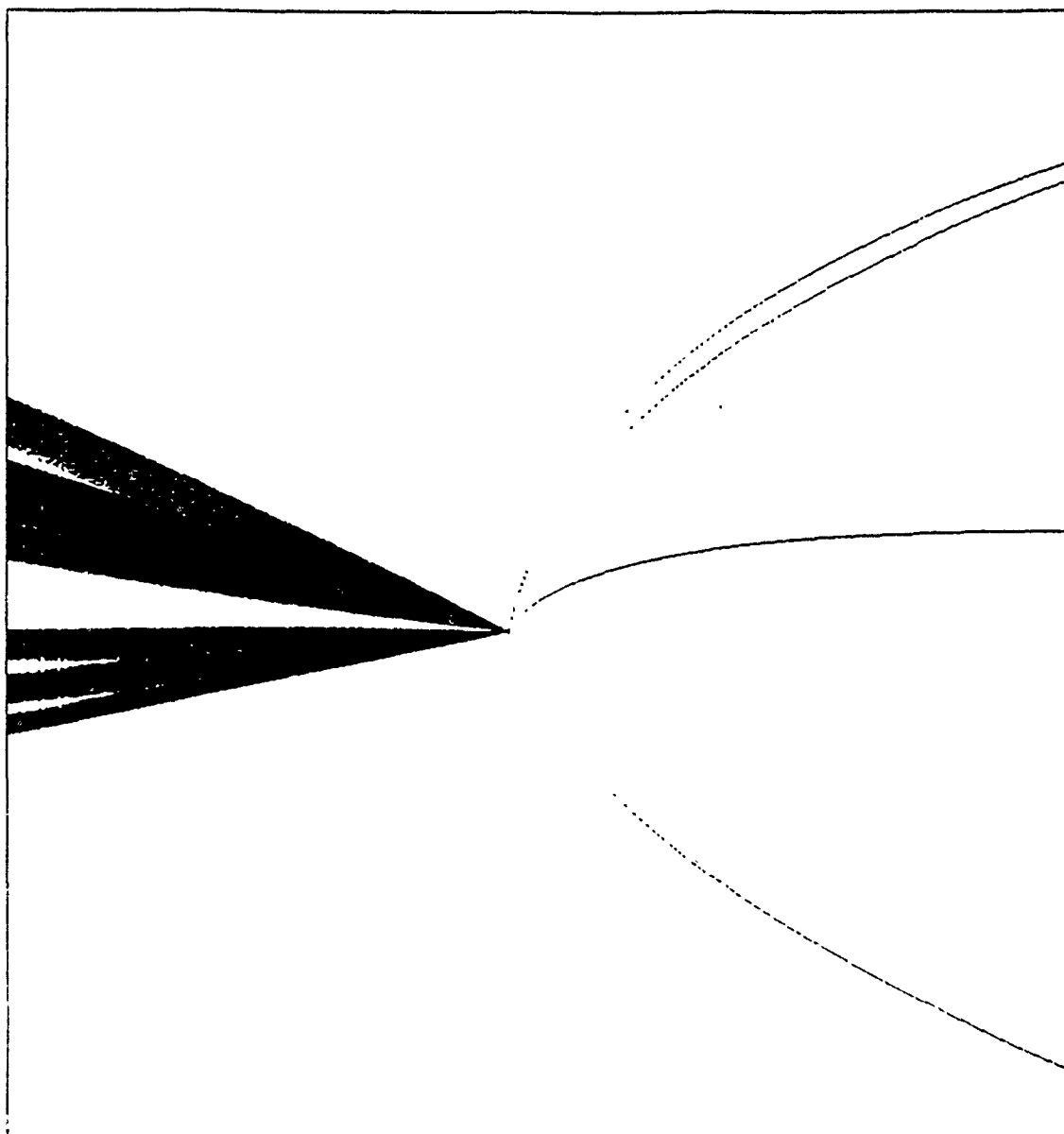


Figure 9b.

Bifurcation diagram of  
 $F_{\mu}(x,y) = (1.4 - x^2 + 0.3y, x)$  if  $x \leq \mu$ , and  
 $= (1.4 + 1.4x + 0.3y - (\mu+1.4)\mu, -4x + 5\mu)$  if  $x \geq \mu$ ,  
 exhibits at  $\mu_0 \approx 0.884$  a border-collision bifurcation from a  
 5-piece chaotic attractor to a period 4 attractor. The parameter  $\mu$   
 (plotted horizontally) varies from 0.874 to 0.895; the coordinate  
 $x$  is plotted vertically,  $0.3 \leq x \leq 1.6$ .

attractor to a 1-piece chaotic attractor, where  $\mu$  (plotted horizontally) varies from 0.87 to 0.895. On both sides of the collision-bifurcation, which occurs at  $\mu_0 \approx 0.884$ , there are infinitely many unstable periodic orbits involved, since the attractors are chaotic. Due to the projection of the picture onto one phase space coordinate the bifurcation diagram seems to show a 2-piece chaotic attractor, but again in phase space one has clearly a 5-piece chaotic attractor.

For the same parameter values, the bifurcation diagram in Figure 9b exhibits a bifurcation from a 5-piece chaotic attractor to a period 4 attractor, where  $\mu$  (plotted horizontally) varies from 0.874 to 0.895. Hence, we may have a border-collision bifurcation from a 5-piece chaotic attractor to a coexisting 1-piece chaotic attractor and a period 4 attractor.

EXAMPLE 8. Now we consider an example in which the curve  $\Gamma_\mu$  is the straight line  $y = -x + \mu$ . In this example we have a moving border. Let the map  $H$  from the plane to itself be defined as above, that is,  $H(x,y) = (A - x^2 + By, x)$ , and define the map  $G_\mu$  ( $-\infty < \mu < \infty$ ) from the plane to itself by

$$G_\mu(x,y) = (A - \mu C - x^2 + Cx + (B+C)y, (B+D)x - Dy - \mu D)$$

The regions  $R_A$  and  $R_B$  are the half planes to the left and the right of the curve  $\Gamma_\mu$ . The map we are investigating is defined being the Henon map on  $R_A$  and the "linear" map  $G_\mu$  on  $R_B$ . Define the one-parameter family of maps  $F_\mu$  from the plane to itself by

$$F_\mu(x,y) = \begin{cases} H(x,y) & \text{if } x \leq -y + \mu \\ G_\mu(x,y) & \text{if } x \geq -y + \mu \end{cases}$$

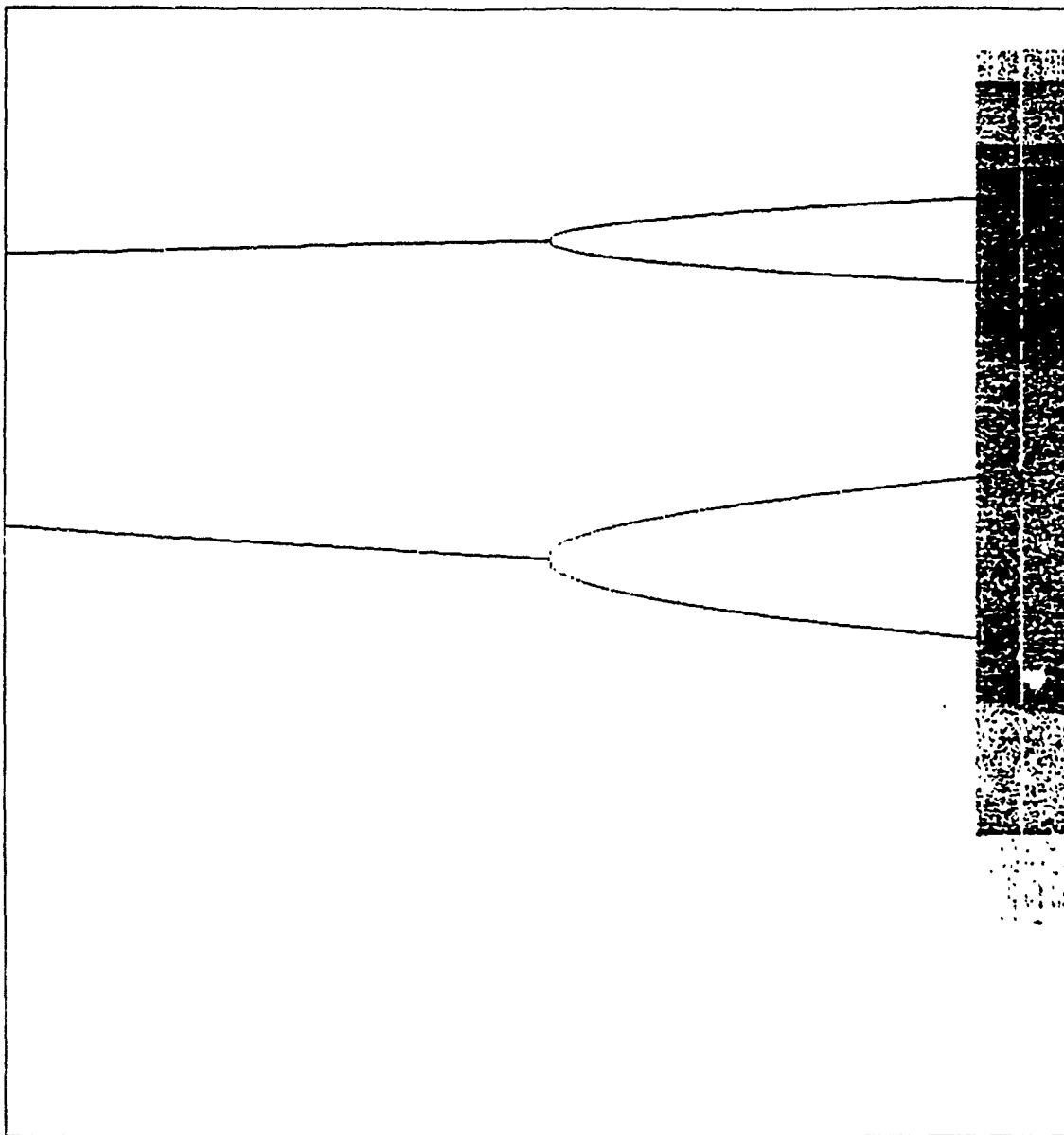


Figure 10a.

Bifurcation diagram of  $F_\mu(x, y) = (1.4 - x^2 + 0.3y, x)$  if  $x \leq -y + \mu$ , and  $F_\mu(x, y) = (1.4 + 0.5\mu - x^2 + 0.5x + 0.2y, -1.3x + y + \mu)$  if  $x \geq -y + \mu$ , exhibits at  $\mu_0 \approx 1.015$  a border-collision bifurcation from a period 4 attractor to a strange chaotic attractor. The parameter  $\mu$  (plotted horizontally) varies from 1.2 to 1; the coordinate  $x$  is plotted vertically,  $-2 \leq x \leq 2$ .

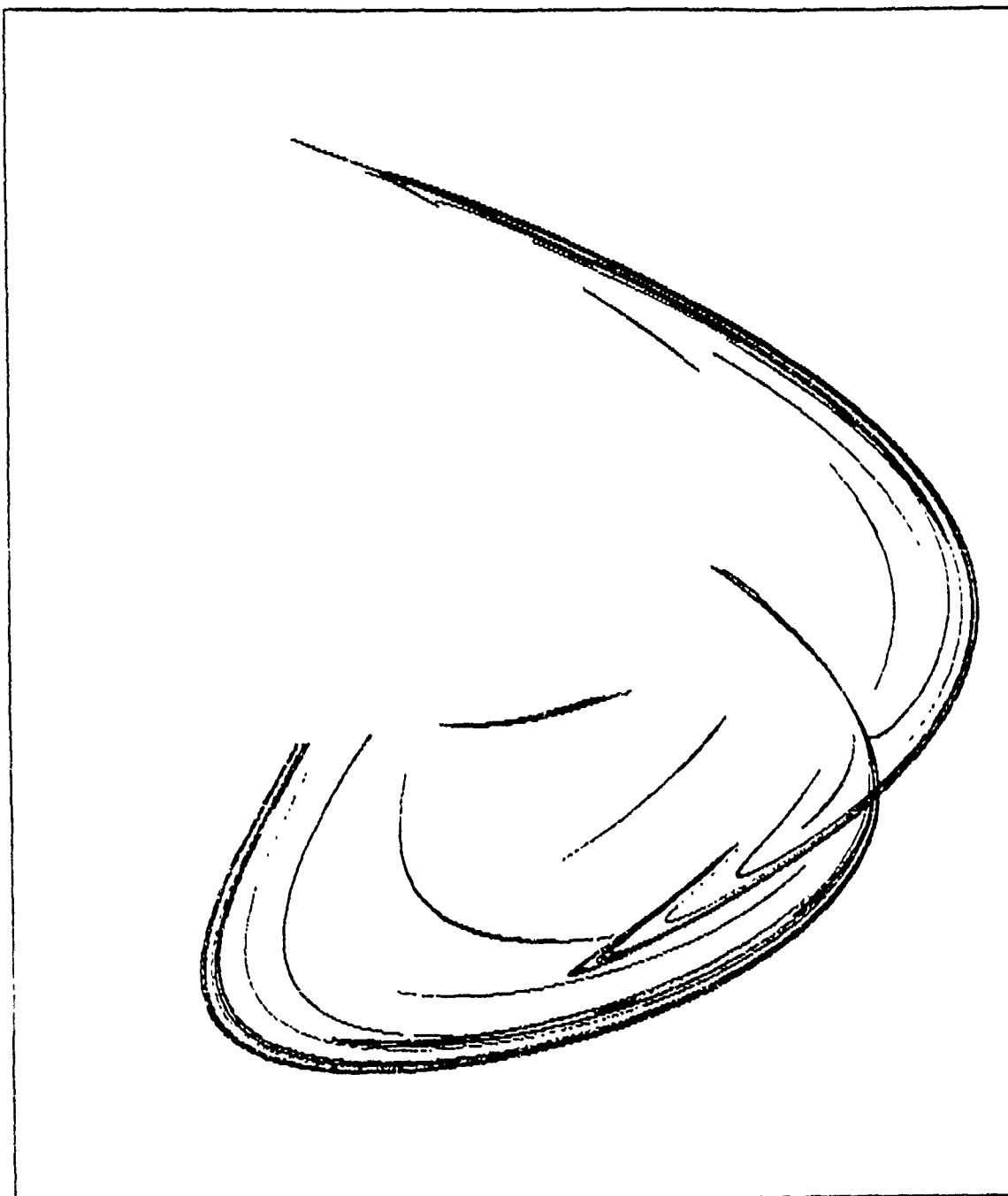


Figure 10b.

The chaotic strange attractor  
 $F_{\mu}(x,y) = (1.4 - x^2 + 0.3y, x)$  if  $x \leq -y + \mu$ , and  
 $= (1.4 + 0.5\mu - x^2 + 0.5x + 0.2y, -1.3x + y + \mu)$  if  
 $x \geq -y + \mu$ , where  $\mu = 1$ . The coordinate  $x$  ( $-2 \leq x \leq 2$ ) is plotted  
horizontally, and the coordinate  $y$  ( $-2 \leq y \leq 2$ ) is plotted  
vertically.

Notice that the map  $F_\mu$  is a piecewise smooth map. We present an example for which the map  $F_\mu$  has a the border-collision bifurcation from a period 4 attractor to a chaotic strange attractor. For  $A = 1.4$ ,  $B = -0.3$ ,  $C = 0.5$ , and  $D = -1$ , the bifurcation diagram in Figure 10a exhibits a bifurcation from a period 4 attractor to a chaotic strange attractor, where  $\mu$  (plotted horizontally) varies from 1.2 to 1. The border-collision bifurcation occurs at  $\mu = \mu_0 \approx 1.015$ . The chaotic strange attractor for  $\mu = 1$  is given in Figure 10b. Hence, we may have a border-collision bifurcation from a period 4 attractor to a chaotic strange attractor.

#### 4. "PERIOD TWO TO PERIOD THREE" BORDER-COLLISION BIFURCATION

In this Section we explain why "period two to period three" border-collision bifurcations occur for two-dimensional piecewise smooth maps. Let  $a$ ,  $b$ ,  $c$ , and  $d$  denote real numbers. Define the one-parameter family  $GL_\mu$  from the plane to itself, by

$$\begin{aligned} GL_\mu(x,y) &= (ax + y, bx) + \mu(1,0) && \text{if } x \leq 0 \\ GL_\mu(x,y) &= (cx + y, dx) + \mu(1,0) && \text{if } x \geq 0 \end{aligned}$$

where  $\mu$  is in an open interval  $I$  including zero. Recall that this family  $GL_\mu$  is equivalent with the piecewise linear map  $P_\mu$ .

Let  $F_\mu$  be a one-parameter family of piecewise smooth maps which has a prototype piecewise linear form at  $\mu = 0$ , and assume that

$$(A1) \quad -a > 1, \quad -c > 1, \quad -d > 1;$$

$$(A2) \quad c^2 + 4d < 0;$$

$$(A3) \quad 0 < a(ac + d) < 1.$$

We want to show that there exists  $\epsilon > 0$  such that if  $|b| < \epsilon$ , then the family  $F_\mu$  has a "period two to period three" border-collision

bifurcation at  $(0,0)$ . First, we show that for  $b = 0$ , the family  $GL_\mu$  has a border-collision bifurcation from a period 2 attractor to a period 3 attractor. We write  $C$  for the set of all one-parameter families of maps  $GL_\mu$  defined above such that  $b = 0$ .

PROPOSITION. At  $\mu = 0$ , every family  $GL_\mu$  in  $C$  has a "period two to period three" border-collision bifurcation at  $(0,0)$ .

PROOF OF THE THEOREM. Assume that the Proposition has been proved. Apply the Proposition and it follows immediately from perturbation results.

The geometrical proof of the Proposition (given below) might give insight why other bifurcations (for example, period 5 to period 2 bifurcation) may occur in piecewise smooth systems. Presumably, the method of proof only works if one of the two maps involved has a zero Jacobian. We first show that a border-collision bifurcation occurs at  $\mu = 0$ , and we present an example to give an idea of the proof.

Let  $GL_\mu$  be in  $C$ . The fixed point  $E_\mu$  of  $F_\mu$  is given by  $E_\mu = (\frac{1}{1-a} \cdot \mu, 0)$  if  $\mu \leq 0$  and  $E_\mu = (\frac{1}{1-c-d} \cdot \mu, \frac{d}{1-c-d} \cdot \mu)$  if  $\mu > 0$ .

In the notation of Section 2, define the matrices  $M_A$  and  $M_B$  by

$$M_A = \begin{bmatrix} a & 1 \\ b & 0 \end{bmatrix}, \quad M_B = \begin{bmatrix} c & 1 \\ d & 0 \end{bmatrix}. \quad \text{The eigenvalues of } M_A \text{ are } 0 \text{ and } a, \text{ so}$$

if  $\mu < 0$  then the fixed point  $E_\mu$  is unstable since  $-a > 1$ . In particular,  $E_\mu$  is a flip saddle if  $\mu < 0$ . The eigenvalues of  $M_B$  are  $0.5c \pm 0.5\sqrt{c^2 + 4d}$  and are complex, since  $c^2 + 4d < 0$ . For  $\mu > 0$  the fixed point  $E_\mu$  is unstable (repelling), since the

product  $-d$  of the eigenvalues of  $M_B$  exceeds 1. The nature of the fixed point  $E_\mu$  is changing from being a flip saddle (in region  $R_A$  which is the left half plane) to a repellor with complex eigenvalues (in region  $R_B$ ) when the parameter  $\mu$  is varied from say  $-0.1$  to  $0.1$ . We conclude that a border-collision bifurcation occurs at  $\mu = 0$  when  $\mu$  is continuously varied from some negative value to a positive value, since the orbit index of  $E_\mu$  changes from 0 to +1. For simplicity of the explanation of this border-collision bifurcation phenomenon, we offer the following example.

EXAMPLE. Consider the one-parameter family  $g_\mu$  from the plane to itself, defined by

$$\begin{aligned} g_\mu(x, y) &= \left(-\frac{5}{4}x + y, 0\right) + \mu \cdot (1, 0) & \text{if } x \leq 0 \\ g_\mu(x, y) &= \left(-2x + y, -\frac{21}{8}x\right) + \mu \cdot (1, 0) & \text{if } x \geq 0 \end{aligned}$$

The bifurcation diagram exhibiting the "period two to period three" bifurcation, is similar to the diagram in figure 1. The family of maps  $g_\mu$  is in the class C, so it is an example for which the result above applies. The idea why a "period two to period three" border-collision bifurcation occurs for the family  $g_\mu$ , is the following.

For  $\mu < 0$ , write  $W_\mu$  for the interval  $[-\frac{1}{5}\mu, \infty) = [-\frac{4}{5}\mu, \infty)$  on the X-axis. We have (1) the image  $g_\mu(p)$  of each point  $p$  on the X-axis but not in  $W_\mu$  is in  $W_\mu$ , and (2) each point  $p$  in  $W_\mu$  is mapped to a point  $p^*$  on the X-axis after two iterates, so  $g_\mu^2(p) = p^*$ . In figure 11, the graph of the corresponding return map  $G$  on  $W_\mu$  which is defined by  $G(x) = g_\mu^2(x, 0)$ , is given. To be more specific,  $G(x) = \frac{25}{16}x - \frac{1}{4}\mu$  for  $\frac{4}{5}\mu \leq x \leq 0$  and  $G(x) = -\frac{1}{8}x - \frac{1}{4}\mu$  for  $x \geq 0$ .

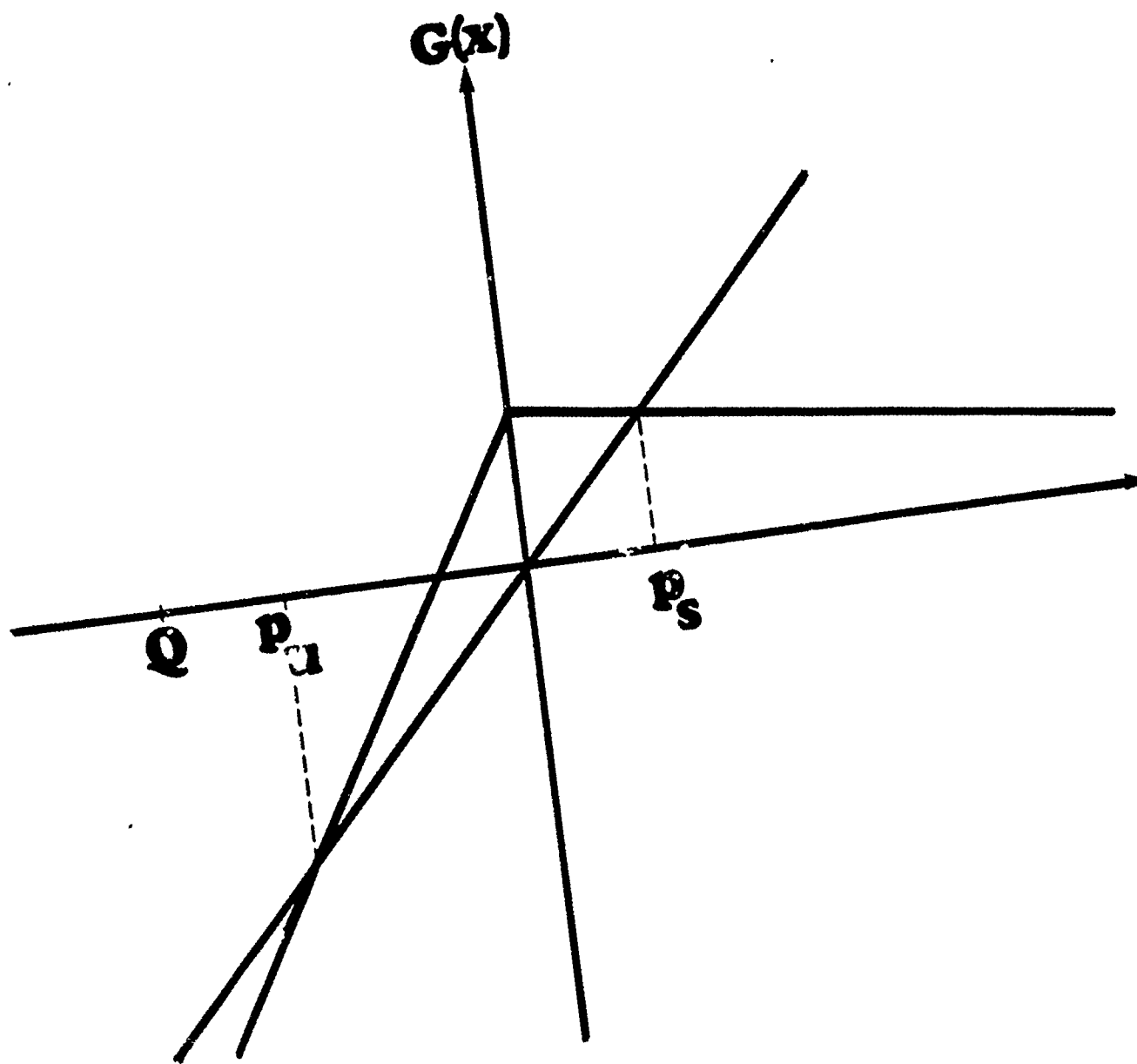


Figure 11.

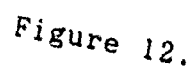
The map  $g_\mu$  is defined by  $g_\mu(x, y) = (-1.25x + y + \mu, 0)$  if  $x \leq 0$ , and  $g_\mu(x, y) = (-2x + y + \mu, -2.625x)$  if  $x > 0$ . For  $\mu < 0$ , the return map  $G$  defined by  $G(x) = g_\mu^2(x, 0)$ , maps the interval  $[0.8\mu, \infty)$  on the  $X$ -axis into the  $X$ -axis. The map  $G$  has an unstable fixed point  $p_u = \frac{4}{9}\mu < 0$  and a stable fixed point  $p_s = -\frac{2}{9}\mu > 0$ .

FIGURE 11

The map  $G$  has two fixed points  $p_u = \frac{4}{9} \cdot \mu < 0$  and  $p_s = -\frac{2}{9} \cdot \mu > 0$ . The fixed point  $p_u$  is unstable since the slope of  $G$  in  $p_u$  is  $\frac{25}{16}$ , and the fixed point  $p_s$  is stable since the slope of  $G$  at  $p_s$  is  $-\frac{1}{8}$ . The properties (1)  $\frac{4}{5} \cdot \mu < p_u = \frac{4}{9} \cdot \mu < 0$ , (2)  $G$  has slope  $\frac{25}{16}$  at  $x$  for  $\frac{4}{5} \cdot \mu < x < 0$ , (3)  $G$  has slope  $-\frac{1}{8}$  for  $x > 0$ , and (4)  $G(0) = -\frac{1}{4} \cdot \mu > 0$ , imply that  $g_\mu$  has a period 2 attractor consisting of the two points  $P_1 = (-\frac{2}{9} \cdot \mu, 0)$  and  $P_2 = g_\mu(P_1) = (\frac{13}{9} \cdot \mu, \frac{7}{12} \cdot \mu)$ . Notice that the norms of both these points converge to zero as  $\mu$  goes to zero, that is, both  $\|P_1\| \rightarrow 0$  and  $\|P_2\| \rightarrow 0$  as  $\mu \rightarrow 0$ . In other words, the period 2 attractor shrinks to a point as  $\mu$  goes to zero; this point to which the period 2 attractor converges is the fixed point of  $g_\mu$  at  $\mu = 0$ . For  $\mu > 0$ , each point  $p$  on the  $X$ -axis is mapped to a point  $p^*$  on the  $X$ -axis after three iterates, so  $g_\mu^3(p) = p^*$ . The graph of the corresponding return map  $H$ , defined by  $H(x) = g_\mu^3(x, 0)$ , is given in figure 12. In particular,  $H(x) = \frac{5}{32} \cdot x - \frac{3}{4} \cdot \mu$  for  $x < 0$ ,  $H(x) = \frac{113}{32} \cdot x - \frac{3}{8} \cdot \mu$  for  $0 \leq x \leq \frac{1}{2} \cdot \mu$ , and  $H(x) = \frac{5}{32} \cdot x + \frac{21}{16} \cdot \mu$  for  $x \geq \frac{1}{2} \cdot \mu$ .

FIGURE 12

The map  $H$  has an unstable fixed point  $p_u = \frac{4}{27} \cdot \mu > 0$  and two stable fixed points  $q_s = -\frac{4}{9} \cdot \mu < 0$  and  $p_s = \frac{14}{9} \cdot \mu > 0$ . Furthermore, for all  $x$  with  $x < p_u$  we have  $\lim_{n \rightarrow \infty} H^n(x) = q_s$ , and for all  $x$  with  $x > p_u$  we have  $\lim_{n \rightarrow \infty} H^n(x) = p_s$ . The properties (1)  $H$  has slope between 0 and 1 for  $x < 0$ , (2)  $H$  has slope bigger than 1 for  $0 < x < \frac{1}{2} \cdot \mu$ , (3)  $H$  has slope between 0 and 1 for  $x > \frac{1}{2} \cdot \mu$ , and (4)  $H(0) = -\frac{3}{8} \cdot \mu < 0$  and  $H(\frac{1}{2} \cdot \mu) = \frac{89}{64} \cdot \mu > \frac{1}{2} \cdot \mu$ , imply  $g_\mu$  has a period 3 attractor consisting of the points  $S_1 = (-\frac{4}{9} \cdot \mu, 0)$ ,  $S_2 = (\frac{14}{9} \cdot \mu, 0)$ , and  $S_3 =$



12. The map  $g_\mu$  is defined by  $g_\mu(x, y) = (-1.25x + y + \mu, 0)$  if  $x \leq 0$ , and  $g_\mu(x, y) = (-2x + y + \mu, -2.625x)$  if  $x > 0$ . For  $\mu > 0$ , the return map  $H$  defined by  $H(x) = g_\mu^3(x, 0)$ , maps the  $X$ -axis into the  $X$ -axis. The map  $H$  has one unstable fixed point  $p_\mu = \frac{4}{27} \cdot \mu > 0$  and two stable fixed points  $q_s = -\frac{4}{9} \cdot \mu < 0$  and  $p_s = \frac{14}{9} \cdot \mu > 0$ .

$(-\frac{19}{9}\mu, -\frac{49}{12}\mu)$ . Notice that the norms of all three points converge to zero as  $\mu$  goes to zero, that is, all three  $\|S_1\| \rightarrow 0$ ,  $\|S_2\| \rightarrow 0$ , and  $\|S_3\| \rightarrow 0$  as  $\mu \rightarrow 0$ . In other words, the period 3 attractor shrinks to a point as  $\mu$  goes to zero; this point to which the period 3 attractor converges is the fixed point of  $g_\mu$  at  $\mu = 0$ . The point  $(\frac{4}{27}\mu, 0)$  is a point of a period 3 orbit which is a regular saddle of the map  $g_\mu$ .

Conclusion: at  $\mu = 0$ , there is a "period two to period three" border-collision bifurcation. END OF THE EXAMPLE.

PROOF OF THE PROPOSITION. Let  $GL_\mu$  be a one-parameter family in the class C, where  $\mu$  is in some interval I. We write  $p_0 = (x_0, y_0)$  for an initial condition and  $p_n = (x_n, y_n)$  for its n-th iterate, that is,  $p_n = GL_\mu^n(p_0)$ , for each  $\mu$ . For the particular initial value  $(0,0)$ , we write  $A_0 = (0,0)$ ,  $A_1 = GL_\mu(A_0)$ ,  $A_2 = GL_\mu(A_1)$ ,  $A_3 = GL_\mu(A_2)$ , and  $A_4 = GL_\mu(A_3)$ .

For each initial value  $p_0 = (x_0, y_0)$  we observe the following fact. If  $x_0 \leq 0$  then  $y_1 = 0$ , and if  $x_0 > 0$  then  $y_1 = dx_0 < 0$ . Hence, it is sufficient to consider initial values in the lower half plane. Hence, from now on, we assume that  $y_0 \leq 0$ .

Assume first,  $\mu < 0$ . Recall that the fixed point  $E_\mu = (\frac{1}{1-a}\mu, 0)$  is unstable, and is a flip saddle, since  $-a > 1$ . Assume that  $p_0 = (x_0, y_0)$  is any initial value with  $y_0 \leq 0$ . Then, if  $x_0 \leq 0$  then  $y_1 = 0$ , and if  $x_0 > 0$ , then  $x_1 = cx_0 + y_0 + \mu < 0$  and so  $y_2 = 0$ . Therefore, it is sufficient to consider points on the X-axis, and we will do so.

Consider the initial value  $p_0 = (0,0) = A_0$ . Computation of the

first four iterates of  $A_0$  yields  $A_1 = (\mu, 0)$ ,  $A_2 = ((a+1)\mu, 0)$ ,  $A_3 = ((c(a+1) + 1)\mu, d(a+1)\mu)$ , and  $A_4 = ((a+1)(ac + d + 1)\mu, 0)$ . The assumptions  $0 < a(ac + d) < 1$  and  $-a > 1$  imply  $-1 < ac + d < 0$  yielding  $0 < x_4 < x_2$ . From  $-1 < ac + d < 0$ , and the assumptions,  $-a > 1$ , and  $-c > 1$  follows that  $c(a + 1) > 0$  and  $d < c$ ; therefore  $|x_3| > |y_3|$ . Hence,  $A_1$  is on the X-axis to the left of  $A_0$ ,  $A_3$  is under and to the left of  $A_1$ , and both  $A_2$  and  $A_4$  are on the X-axis to the right of  $A_0$  and  $A_4$  is between  $A_0$  and  $A_2$ .

First we consider the image of the X-axis. Let  $p_0 = (x_0, y_0)$  be any point on the X-axis. The image of the right half of the X-axis with end point  $A_0$  is the half line through  $A_3$  with end point  $A_1 = GL_\mu(A_0)$ , since  $p_1 = (cx_0 + \mu, dx_0)$  for  $x_0 > 0$ . The image of the left half of the X-axis with end point  $A_0$  is the half line on the X-axis to the right of  $A_1$  with end point  $A_1$ , since  $p_1 = (ax_0 + \mu, 0)$  for  $x_0 \leq 0$ .

Define  $Q = (-\frac{1}{a} \cdot \mu, 0) = (x_Q, 0)$  and  $R = (\frac{a^2}{(1-a)(ac + d)} \cdot \mu, 0) = (x_R, 0)$ . The point  $Q$  is mapped to  $A_0$  iterating  $GL_\mu$  once, that is,  $GL_\mu(Q) = A_0$ , and  $Q$  is on the X-axis between  $A_1$  and  $E_\mu$  since  $A_1 = (\mu, 0)$ ,  $E_\mu = (\frac{1}{1-a} \cdot \mu, 0)$  and  $-a > 1$ . The point  $R$  is on the X-axis to the right of  $A_0$ , and  $R$  is mapped to  $E_\mu$  iterating  $GL_\mu$  twice, that is,  $GL_\mu^2(R) = E_\mu$ .

Let  $p_0 = (x_0, 0)$  be any point. Straightforward computation gives the following. If  $x_0 > 0$  (that is,  $p_0$  is on the X-axis to the right of  $A_0$ ) then  $p_1 = (cx_0 + \mu, dx_0)$  and  $p_2 = ([ac + d]x_0 + (1+a)\mu, 0)$ , so  $p_2$  is on the X-axis. If  $x_0 = 0$  (that is,  $p_0 = A_0$ ) then  $p_1 = (\mu, 0)$  and  $p_2 = ((1+a)\mu, 0)$ , so  $p_2$  is on the X-axis to the right of  $A_0$ . If  $-\frac{1}{a} \cdot \mu \leq x_0 < 0$  (that is,  $p_0$  is on the X-axis between  $Q$  and  $A_0$ ) then  $p_1 = (ax_0 + \mu, 0)$  and  $p_2 = (a(ax_0 + \mu) + \mu,$

0), so  $p_2$  is on the X-axis. If  $x_0 < -\frac{1}{a}\mu$  (that is,  $p_0$  is to the left of Q) then  $p_1 = (ax_0 + \mu, 0)$  and  $p_2 = (c(ax_0 + \mu) + \mu, d(ax_0 + \mu))$  and  $p_3 = ([a^2c + ad]x_0 + (ad + a + d + 1)\mu, 0)$ , and so  $p_3$  is on the X-axis while  $p_2$  is not. Summarizing, for each point  $p_0$  on the X-axis to the right of Q we have  $p_2 = GL_\mu^2(p_0)$  is on the X-axis. Therefore, we have a return map on the interval consisting of the points on the X-axis to the right of Q.

Let  $G$  denote the return map of  $GL_\mu$  on  $[Q, \infty)$ , so  $G(x) = GL_\mu^2(x, 0)$  for each  $x \geq x_Q$ . The above results imply  $G(x) = a^2x + (1+a)\mu$  for  $-\frac{1}{a}\mu \leq x \leq 0$ , and  $G(x) = (ac+d)x + (1+a)\mu$  for  $x \geq 0$ . The graph of  $G$  is similar to figure 11. The map  $G$  has two fixed points, namely  $p_u = \frac{1}{1-a}\mu$ , and  $p_s = \frac{a+1}{1-ac-d}\mu$ , and  $p_u < 0 < p_s$ . The fixed point  $p_u$  is unstable since the slope of  $G$  in  $p_u$  is  $a^2 > 1$ , and the fixed point  $p_s$  is stable since the slope of  $G$  at  $p_s$  is  $ac+d$  for which  $-1 < ac+d < 0$ . Furthermore, for all  $x$  with  $p_u < x < x_R$  we have  $\lim_{n \rightarrow \infty} G^n(x) = p_s$ . The properties (1)  $x_Q < \frac{1}{1-a}\mu < 0$ , (2)  $G$  has slope  $a^2 > 1$  if  $x_Q < x < 0$ , (3)  $G$  has slope  $-1 < ac+d < 0$  for  $x > 0$ , and (4)  $G(0) > 0$ , imply that  $GL_\mu$  has a period 2 attractor consisting of the points  $P_1 = (\frac{a+1}{1-ac-d}\mu, 0)$  and  $P_2 = GL_\mu(P_1) = (\frac{c-d+1}{1-ac-d}\mu, \frac{(a+1)d}{1-ac-d}\mu)$ . Notice that the norms of both these points converge to zero as  $\mu$  goes to zero, that is, both  $\|P_1\| \rightarrow 0$  and  $\|P_2\| \rightarrow 0$  as  $\mu \rightarrow 0$ . Hence, the period 2 attractor shrinks to a point as  $\mu$  goes to zero; this point to which the period 2 attractor converges is the fixed point of  $GL_\mu$  at  $\mu = 0$ .

Now assume  $\mu = 0$ . Assume  $p_0 = (x_0, y_0)$  is any initial value with  $y_0 \leq 0$ , then  $x_0 \leq 0$  implies  $y_1 = 0$ , and  $x_0 > 0$  implies  $x_1 = cx_0$  yielding  $y_2 = 0$ . Hence, it is sufficient to consider points on the X-axis. Let  $p_0 = (x_0, 0)$  be given. If  $x_0 < 0$  then  $p_1 = (ax_0, 0)$

which is on the positive X-axis. If  $x_0 = 0$  then  $p_1 = (ax_0, 0)$  and so  $p_0$  is the fixed point of  $GL_0$ . If  $x_0 > 0$ , then  $p_1 = (cx_0, dx_0)$ , and  $p_2 = ((ac+d)x_0, 0)$ . Consequently the point  $A_0 = (0,0)$  is a globally stable fixed point of  $GL_0$ , since  $-1 < ac + d < 0$ .

Now assume  $\mu > 0$ . The fixed point  $E_\mu = (\frac{1}{1-c-d}\cdot\mu, \frac{d}{1-c-d}\cdot\mu)$  is unstable with complex eigenvalues since it was assumed  $-d > 1$  and  $c^2 + 4d < 0$ . Assume  $p_0 = (x_0, y_0)$  is any initial value with  $y_0 \leq 0$ . Then  $x_0 \leq 0$  implies  $y_1 = 0$ , and if  $x_0 \geq -\frac{1}{c}\cdot\mu$  then  $x_1 = cx_0 + y_0 + \mu \leq 0$  and so  $y_2 = 0$ . If  $0 < x_0 < -\frac{1}{c}\cdot\mu$  then  $x_1 = cx_0 + y_0 + \mu$  and  $y_1 = dx_0 < 0$ ; hence, if  $x_1 \leq 0$  then  $y_2 = 0$ , else if  $x_1 > 0$  (and so  $0 \leq -y_0 < \mu + cx_0$ ),  $x_2 = cx_1 + y_1 + \mu = c^2x_0 + cy_0 + c\mu + dx_0 + \mu < (c^2 - c(\mu + cx_0) + c\mu + dx_0 + \mu = dx_0 + \mu < 0$ , and so  $y_3 = 0$ . Therefore, it is sufficient to consider points on the X-axis.

Let  $p_0 = (x_0, y_0) = (x_0, 0)$  be any point on the X-axis. If  $x_0 \leq 0$  then  $p_1 = GL_\mu(p_0) = (ax_0 + \mu, 0) = (x_1, y_1)$ , so  $x_1 > 0$ . Every point  $q_0 = (w_0, 0)$  such that  $w_0 < x_0 \leq 0$  satisfies  $q_1 = GL_\mu(q_0) = (aw_0 + \mu, 0) = (w_1, z_1)$ , so  $w_1 > x_1 > 0$ . The conclusion is that points on the X-axis to the left of  $A_0 = (0,0)$  are mapped monotonically into the X-axis to the right of  $(\mu, 0)$ .

Let  $p_0 = (0,0)$ . A simple computation shows  $p_1 = (\mu, 0)$ ,  $p_2 = ((1+c)\mu, d\mu)$ ,  $p_3 = ((ac + a + d + 1)\mu, 0)$ , and  $p_4 = (ax_3 + \mu, 0)$ . Notice  $x_3 < 0$ , hence  $x_4 > \mu = x_1$ . Recall that  $p_0 = A_0$ ,  $p_1 = A_1$ ,  $p_2 = A_2$ ,  $p_3 = A_3$ , and  $p_4 = A_4$ . The conclusion is that  $A_0$ ,  $A_1$ ,  $A_3$ , and  $A_4$  are on the X-axis, and  $A_3$  is to the left of  $A_0$ , and both  $A_1$  and  $A_4$  are to the right of  $A_0$  with  $A_1$  between  $A_0$  and  $A_4$ .

Let  $p_0 = (x_0, 0)$  be any point on the X-axis for which  $x_0 > 0$ . Then  $p_1 = (cx_0 + \mu, dx_0)$ . Notice that if  $x_0 = -\frac{1}{c}\cdot\mu$  then  $x_1 = 0$  and  $y_1 = -\frac{d}{c}\cdot\mu$ . Write  $B_0 = (-\frac{1}{c}\cdot\mu, 0)$ ,  $B_1 = GL_\mu(B_0)$ ,  $B_2 = GL_\mu(B_1)$ , and

$B_3 = GL_\mu(B_2)$ . Then  $B_1 = (0, -\frac{d}{c}\mu)$ ,  $B_2 = ((1-\frac{d}{c})\mu, 0)$ , and  $B_3 = ([a(1-\frac{d}{c}) + 1]\mu, 0)$ . Notice that  $B_1$  denotes the point on the Y-axis at which the line segment  $[A_1, A_2]$  intersects the Y-axis, and that  $B_2$  is a point on the X-axis to the left of  $A_0$ . The assumptions  $-a > 1$  and  $0 < a(ac + d) < 1$  imply  $ac + d < 0$  and we obtain that the point  $A_3 = (ac + a + d + 1)\mu, 0$  is on the X-axis to the left of  $B_2$ .

The image of the half line  $[A_1, \infty)$  through  $A_2$  under the map  $GL_\mu$  is the kinked half line  $[A_2, B_2] \cup [B_2, \infty)$  through  $A_3$ . The image of this kinked half line is on the X-axis. In particular, the image of the half line  $[B_2, \infty)$  through  $A_3$  is  $[B_3, \infty)$  on the X-axis to the right of  $A_1 = (\mu, 0)$ , and the image of the line segment  $[A_2, B_2]$  is  $[A_3, B_3]$ .

Let  $p_0 = (x_0, 0)$  be any point on the X-axis. Straightforward computation shows the following. If  $x_0 \geq -\frac{1}{c}\mu$  (that is,  $p_0$  is to the right of  $B_0$ ) then  $p_1 = (cx_0 + \mu, dx_0)$ ,  $p_2 = ([ac + d]x_0 + (a+1)\mu, 0)$ , and  $p_3 = (a[ac+d]x_0 + [a(a+1) + 1]\mu, 0)$ . Hence, both  $p_2$  and  $p_3$  are on the X-axis for  $x_0 \geq -\frac{1}{c}\mu$ . If  $0 \leq x_0 \leq -\frac{1}{c}\mu$  (that is,  $p_0$  is on the X-axis between  $A_0$  and  $B_0$ ) then  $p_1 = (cx_0 + \mu, dx_0)$ ,  $p_2 = ([c^2 + d]x_0 + (c+1)\mu, cdx_0 + d\mu)$ , and since  $(c^2 + d)x_0 + (c+1)\mu < -c\mu - \frac{d}{c}\mu + c\mu + \mu = (1 - \frac{d}{c})\mu < 0$ , we have  $p_3 = ([ac^2 + ad + cd]x_0 + [ac + a + d + 1]\mu, 0)$ , so the point  $p_3$  is on the X-axis. If  $x_0 < 0$  then  $p_1 = (ax_0 + \mu, 0)$ ,  $p_2 = (acx_0 + (c+1)\mu, adx_0 + d\mu)$ , and  $p_3 = (a[ac + d]x_0 + [ac + a + d + 1]\mu, 0)$ , so the point  $p_3$  is on the X-axis. The conclusion is that for each point  $p_0 = (x_0, 0)$  on the X-axis, the third iterate of  $p_0$  is also on the X-axis, that is,  $GL_\mu^3(p_0) = (x_3, 0)$ . Hence, a return map of  $GL_\mu$  exists on the X-axis. We call this return map  $H$ , so  $H(x) =$

$GL_\mu^3(x,0)$ . The above results imply

$$H(x) = (a^2c + ad)x + (ac + a + d + 1) \cdot \mu \text{ for } x < 0,$$

$$H(x) = (ac^2 + ad + cd)x + (ac + a + d + 1) \cdot \mu \text{ for } 0 \leq x \leq -\frac{1}{c} \cdot \mu, \text{ and}$$

$$H(x) = (a^2c + ad)x + (a^2 + a + 1) \cdot \mu \text{ for } x_0 \geq -\frac{1}{c} \cdot \mu. \text{ The graph of } H$$

is similar to figure 12. The map  $H$  has three fixed points, namely

$$q_s = \frac{ac + a + d + 1}{1 - a(ac+d)} \cdot \mu < 0, \quad p_u = \frac{-(ac + a + d + 1)}{c(ac+d) + ad - 1} \cdot \mu, \text{ and } p_s = \frac{a^2 + a + 1}{1 - a(ac+d)} \cdot \mu > 0. \text{ The fixed point } p_u \text{ is unstable since the slope}$$

$ac^2 + ad + cd$  of  $H$  in  $p_u$  is bigger than 1, and the two fixed points  $q_s$  and  $p_s$  is stable since the slope  $a^2c + ad$  of  $H$  at both  $q_s$  and

$p_s$  is between 0 and 1. Furthermore, for all  $x$  with  $x < p_u$  we have

$$\lim_{n \rightarrow \infty} H^n(x) = q_s, \text{ and for all } x \text{ with } x > p_u \text{ we have } \lim_{n \rightarrow \infty} H^n(x) = p_s.$$

The properties (1)  $H$  has slope between 0 and 1 for  $x < 0$ ,

(2)  $H$  has slope bigger than 1 for  $0 < x < -\frac{1}{c} \cdot \mu$ , (3)  $H$  has slope

between 0 and 1 for  $x > -\frac{1}{c} \cdot \mu$ , and (4)  $H(0) < 0$  and  $H(-\frac{1}{c} \cdot \mu) > -\frac{1}{c} \cdot \mu$ ,

imply that  $GL_\mu$  has a period 3 attractor consisting of the points

$$S_1 = \left( \frac{ac + a + d + 1}{1 - a(ac+d)} \cdot \mu, 0 \right), \quad S_2 = \left( \frac{a^2 + a + 1}{1 - a(ac+d)} \cdot \mu, 0 \right), \text{ and}$$

$$S_3 = \left( \left[ c \cdot \frac{a^2 + a + 1}{1 - a(ac+d)} + 1 \right] \cdot \mu, d \cdot \frac{a^2 + a + 1}{1 - a(ac+d)} \cdot \mu \right). \text{ Notice that the}$$

norms of all three points converge to zero as  $\mu$  goes to zero, that

is, all three  $\|S_1\| \rightarrow 0$ ,  $\|S_2\| \rightarrow 0$ , and  $\|S_3\| \rightarrow 0$  as  $\mu \rightarrow 0$ . Hence,

the period 3 attractor shrinks to a point as  $\mu$  goes to zero; this

point to which the period 3 attractor converges is the fixed point

of  $GL_\mu$  at  $\mu = 0$ .

The point  $(p_u, 0)$  is a point of a period 3 orbit which is a regular saddle of the map  $GL_\mu$ . We conclude: at  $\mu = 0$ , there is a "period two to period three" border-collision bifurcation. This completes the proof of the Proposition.

## 5. DISCUSSION AND CONCLUDING REMARKS.

We have presented bifurcation phenomena, which we call "border-collision bifurcations". These bifurcations occur when the nature of a fixed point (or periodic point) of a piecewise smooth system changes when it collides with the border of two regions. An interesting case occurs when the fixed point changes from being a flip saddle to a repellor with complex eigenvalues, at the parameter value where it collides with the border of two regions. We have presented a variety of examples based on the piecewise linear map  $P_\mu$  and the Henon map. In particular, we have shown the occurrence of a "period two to period three" border-collision bifurcation for maps in the class C.

We point out that the border-collision bifurcation can be expected to occur in many piecewise smooth models. In particular, the "period two to period three" bifurcation phenomenon can be expected to occur in many linear models with constraints.

Assume for the piecewise linear map  $P_\mu$  that the fixed point  $E_\mu$  is a flip saddle in the left half plane and a repellor with complex eigenvalues in the right half plane.

QUESTION 1. Does there exist a classification of the border-collision bifurcations for  $P_\mu$  in the case where a period 2 attractor converges to the fixed point  $(0,0)$  when  $\mu$  goes to 0?

QUESTION 2. More generally, is it possible to give a classification of the border-collision bifurcations for the piecewise linear map  $P_\mu$ ?

QUESTION 3. When the plane is subdivided in  $N$  regions, where  $N$

is at least 3, do there exist border-collision bifurcations that do not occur when there are only 2 regions, and in particular bifurcations that persist despite small perturbations?

#### REFERENCES

[AYY] K.T. Alligood, E.D. Yorke and J.A. Yorke (1987). Why period-doubling cascades occur: period orbit creation followed by stability shedding. *Physica* 28D, 197-205.

[GH] J. Guckenheimer and P. Holmes (1983). Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Applied Mathematical Sciences 42, Springer-Verlag, New-York, etc.

[GOPY] C. Grebogi, E. Ott, S. Pelikan, and J.A. Yorke (1984). Strange attractors that are not chaotic", *Physica D* 11, 261-268.

[HN] C.H. Hommes and H.E. Nusse (1990). "Period three to period two" bifurcation for piecewise linear models. Submitted for publication

[HNS] C.H. Hommes, H.E. Nusse, and A. Simonovits (1990). Hicksian cycles and chaos in a socialist economy. Research memorandum 382, Institute of Economic Research, University of Groningen

[K] H.B. Keller (1987). Numerical methods in bifurcation problems. Springer-Verlag, Berlin, etc.

[MY] J. Mallet-Paret and J.A. Yorke (1982). Snakes: oriented families of periodic orbits, their sources, sinks and continuation. *J. Differential equations* 43, 419-450

[NY] H.E. Nusse and J.A. Yorke (1989). A procedure for finding

numerical trajectories on chaotic saddles. Physica 36D, 137-156.

[R] D. Ruelle (1989). Elements of Differentiable Dynamics and Bifurcation Theory. Academic Press, Inc., San Diego and London

[S] R. Seydel (1988). From Equilibrium to Chaos. Practical Bifurcation and Stability Analysis. Elsevier Science Publ. Co., Inc. New York, Amsterdam and London

[Y] J.A. Yorke (1990). DYNAMICS: An interactive program for IBM PC clones.